Université de Strasbourg

ENGEES

Image Ville Environnement

ED 413

Ecole Doctorale des Sciences de la Terre et de l'Environnement

Laboratoire Image Ville et Environnement LIVE UMR 7362

THÈSE

présentée par :

Corinne GRAC

Soutenue le :

19 décembre 2019

Pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline : Géochimie

Fouille temporelle des indicateurs physico-chimiques et biologiques pour l'évaluation de l'état, des pressions et de la capacité de résilience des rivières

Rapporteurs

Philippe USSEGLIO-POLATERA	Professeur, Université de Lorraine, Metz
Yorick REYJOL	Chef d'équipe UMS Patrinat, MNHN, AFB, CNRS, Paris

Examinatrices

Gabrielle THIÉBAUT	Professeure, Université de Rennes
Maguelonne TEISSEIRE	Directrice de recherche, IRSTEA, Montpellier

Directeurs de thèse

Jean-Nicolas BEISEL	Professeur, ENGEES, Strasbourg
Michèle TRÉMOLIÈRES	Professeure émérite, Université de Strasbourg

Membre invitée

Florence LE BER	Ingénieure Générale des Ponts, des Eaux et des Forêts, ENGEES, Strasbourg

A Yannick, Ulysse et Capucine,

Mon chemin, ma fierté, mon réconfort.

« Et au milieu, coule une rivière… »

RETROSPECTIVE & REMERCIEMENTS

J'appartiens à la catégorie des besogneuses, besogneuse mais chanceuse, notamment dans les rencontres qui ont jalonnées mon parcours professionnel et m'ont permis d'aboutir à ce projet de thèse et à sa réalisation ; rencontres que je veux remercier ici.

Ma carrière a commencé sous les tropiques à la production de données biologiques. Je reconnais parfois encore annoncer mi- fanfaronnant, mi- plaisantant « le premier réseau de surveillance des rivières de La Réunion, c'est moi ; le premier réseau de surveillance des rivières de Guadeloupe, c'est moi !». C'est vrai et j'en conserve une certaine fierté. Un guide des invertébrés de La Réunion est encore là pour l'attester. Certes, depuis, de nombreux travaux successifs ont permis aux départements d'outre-mer de se doter d'indicateurs biologiques adaptés. Mais, début 2000, à la DIREN Guadeloupe, avec mon chef de service, Eric Muller, nous réfléchissions déjà à structurer les données produites et nous dégagions des crédits pour créer, avec une informaticienne, la première base de données locale, au format SANDRE.

Et puis, en 2003, pour des raisons personnelles, je choisis l'exotisme de l'Alsace, et grâce à Daniel Loudière, alors directeur de l'ENGEES, j'entre alors dans une catégorie, qui se fait à présent rare: un ingénieur agriculture et environnement (IAE), occupant un poste d'enseignant-chercheur. A l'ENGEES, je partage un bureau avec Florence Le Ber, informaticienne, nouvellement arrivée comme moi. Depuis, avec Florence, nous collaborons autour des données, des bases de données, des méthodes de fouille, objet de ses recherches. Mon chemin n'aurait pas pu se faire sans Florence, à la fois partenaire de recherche, porteuse du projet FresQueau, qui nous a ouvertes à d'autres collaborations, mais aussi, à présent ma supérieure hiérarchique me soutenant et m'ayant permis, avec mon directeur actuel, Jean-François Quéré, de mener ce projet de thèse pendant mon activité. C'est grâce aussi à Florence et Jean-François qu'en 2013 l'effectif d'hydroécologue de l'ENGEES a doublé avec l'arrivée d'un professeur : Jean-Nicolas Beisel ; et également grâce à eux et Dominique Badariotti, directeur du LIVE, et Laurent Schmitt, à l'époque directeur-adjoint du LIVE, que nous avons pu rejoindre ce laboratoire de recherche.

Jean-Nicolas non seulement a cru à ce projet ambitieux de thèse, dont je parlais sans trop y croire, mais encore m'a permis de poser le sujet et a accepté d'être mon Directeur de thèse. Avec ma co-directrice de thèse, Michèle Trémolières,

ma collègue, mon autre partenaire de recherche et mon mentor en écologie alluviale rhénane depuis mon arrivée en Alsace, ils ont eu la patience de m'écouter, de m'aider à structurer mes idées, de me soutenir, de m'accompagner, parfois de m'aiguillonner, bref de me diriger. Là encore Florence a toujours été présente dans cet encadrement.

Aussi, merci à vous Florence, Michèle, Jean-Nicolas, Jean-François, Dominique, et Laurent. Merci également à Thierry Leviandier, mon premier chef de service à l'ENGEES et le premier à avoir évoqué l'idée d'une thèse.

Ce projet ne serait pas sans quatre autres personnes. D'abord deux informaticiens : Agnès Braud, collègue de Florence, associée à nos projets depuis 2003, et Xavier Dolques. Xavier a patiemment construit PRESTOR, pour qu'il s'adapte à la fois aux données des rivières et à mes attentes. Il a produit des milliers de motifs. Il m'a aidé à les trier et à les interpréter. C'est grâce à votre patience, à tous les deux, votre écoute, votre travail, vos conseils et ceux de Florence que j'ai pu avancer dans ce monde de la fouille de données. La collaboration née de FresQueau, m'a permis également de rencontrer Maguelonne Teisseire, informaticienne à l'origine des motifs avec Mickaël Fabrègue, et Flavie Cernesson, hydrologue, avec qui nous avons notamment mené la collecte, la structuration des données et muri les questions des « thématiciens » à soumettre aux fouilleurs. Nos échanges qui continuent me permettent toujours de progresser dans la formalisation des questions et des solutions à y apporter. Merci donc à vous quatre.

A l'origine, cette thèse devait intégrer une partie de résultats produits avec des analyses multivariées, mais pour diverses raisons cela n'a pas été possible. Je tiens à remercier Photis Nobelis, statisticien retraité, qui a patiemment et amicalement passé des heures avec moi pour m'initier à la subtilité statistique et à la programmation sous R.

Merci aussi à mon collègue Olivier Schlumberger qui a repris, pendant deux ans, une partie de mes enseignements pour me dégager du temps à ce projet de thèse.

Depuis plusieurs années, cette thèse occupe une partie des conversations avec mes collègues. Merci à Agnès Herrmann, Véronique Brid, Caroline Lienhard, pour votre soutien inconditionnel et votre aide. Merci à Estelle Baehrel, Caty Werey, François Destandau, Eliane Propeck, José Vasquez, Denis Cassard, Marie-Pierre Ottermatte, David Eschbach, Laurent Hardion et Isabelle Combroux pour votre écoute et vos conseils. Merci à Juliane Wiederkehr, Nadia Fernandez, Cybill

Staenzel et Albin Meyer, qui ont été mes étudiants, puis ont inversé les rôles en me soutenant et en m'aidant.

Merci également au directeur de l'école doctorale, Jérôme vand der Woerd, qui s'inquiète régulièrement de mon avancement et à son assistante Fayza Fallah, pour son aide dans mes démarches administratives.

La recherche ne se fait pas sans moyens, je tiens également à remercier les personnes qui ont cru aux projets proposés au sein de la DREAL Alsace, l'Agence de l'Eau Rhin-Meuse, l'ONEMA puis l'AFB.

Enfin, en fin de parcours de thèse, il faut un jury ! Merci à Philippe Usseglio-Polatera et Yorick Reyjol d'avoir accepté d'être les rapporteurs de ce travail. Merci à Gabrielle Thiébaut et Maguelonne Teisseire d'avoir accepté d'en être les examinatrices.

Un chemin professionnel ne va pas sans un chemin personnel, … sans ma raison personnelle qui m'a conduite en Alsace : Yannick et ceux que nous y avons construits : nos enfants Ulysse, Capucine. Cette thèse a pris une certaine place dans notre vie quotidienne, elle a parfois monopolisé nos conversations. Elle a inversé les rôles et amené nos enfants à encourager leur maman du haut de leurs 8 et 10 ans. Sans vous, votre soutien permanent à tous les trois, votre compréhension, vos encouragements, ce chemin aurait été impossible. Les repères de temps des parents sont généralement les niveaux de classes de leurs enfants : ma première inscription en doctorat date du CM2 de mon fils, il vient d'entrer au lycée. Il est temps de finir!

Merci aussi pour votre écoute et votre soutien à mes parents, Simone et Charly, à mes frères : Michel et Nicolas, à mes belles-sœurs Audrey, Camille (« *Don't get it, get it done* ») et Anne, à mes beaux-parents, dont Marlyse, toujours dévouée, à mes amies : Béa, Vanessa, Sylvaine, Edith, Marie-Pierre et à mes coachs personnels : Franck et puis Richard. Et toujours une pensée pour notre Eric. Que m'aurait-il dit? Avec son large sourire bienveillant : « Engage tes skis dans la pente ! Tu vas voir : ça va être gaz, mais ça va aller ma sœur ! »

PREAMBULE

Les travaux présentés ici s'inscrivent principalement dans trois projets de recherche successifs ayant tous pour objet l'étude de l'état des écosystèmes rivières.

Le **projet INDICES** (financement Agence de l'Eau Rhin-Meuse & DIREN – Direction Régionale de l'Environnement – Alsace ; porteuse Corinne Grac ; 2005-2010) a initié la collaboration entre thématiciens hydroécologues et informaticiens spécialistes en fouilles de données. Nous avons acquis les données physico-chimiques et de cinq groupes biologiques sur 40 stations en plaine d'Alsace, nous les avons structurées dans une base de données au format SANDRE, nous avons utilisé la méthode de fouille des treillis de Gallois et proposer une méthode de diagnostic multi-indices biologiques de l'état des rivières.

Le **projet FresQueau** (financement ANR – Agence Française de la Recherche – Modèles Numériques 2011 ; porteuse Florence Le Ber ; co-responsables des tâches acquisition des données et formulation des questions des thématiciens : Corinne Grac & Flavie Cernesson ; 2011-2015), a permis de développer et d'appliquer différentes méthodes de fouille de données à la recherche de relations pressions-impacts et à l'évaluation de la qualité écologique des cours d'eau au sens de la Directive Cadre européenne sur l'Eau de 2000 (DCE), sur les données 2000-2010 de deux bassins (Rhin-Meuse et Rhône-Méditerranée-Corse). Parmi ces méthodes de fouille, Fabrègue et al. (2014) ont développé un algorithme spécifique d'extraction de motifs temporels fermés partiellement ordonnés, dits « motifs ».

Le **projet d'extension nationale de FresQueau** (financement ONEMA – Office National des Milieux Aquatiques- ; porteuse Corinne Grac ; 2015-2017) a permis : 1) d'étendre l'extraction des motifs à la base de données de surveillance nationale des rivières contenant les données physico-chimiques et biologiques de 2007 à 2013, 2) de développer une plateforme PRESTOR sous licence libre, permettant à un opérateur d'extraire ces motifs en pouvant faire varier plusieurs critères de choix.

LISTE DES VALORISATIONS SCIENTIFIQUES REALISEES DANS LE CADRE DE CE TRAVAIL

ARTICLE EN SOUMISSION

Grac, C., Cernesson, F., Dolques, X., Braud, A., Herrmann A., Labat, F., Teisseire, M., Trémolières, M., Le Ber, F., 2019. Which data mining method to use for the evaluation of river ecological status – Feedback on a close collaboration between data scientists and hydro-scientists. *Enviromental Modelling and Software. Soumis le 22 août 2019.*

(3ème soumission après une 1ère fois à *Sciences of Total Environnement* le 25 mai 2017, refusé le 22 juin 2017 car ne correspondant pas aux attentes de la revue ; puis une 2ème fois à *Plos One* le 6 octobre 2017, refusé le 2 août 2018, suite à la demande de révisions majeures par les relecteurs et « malgré son intérêt et l'important travail réalisé »)

ARTICLE EN REVISION POUR UNE NOUVELLE SOUMISSION :

Grac, C., Dolques, X., Braud, A., Trémolières M., Beisel, J-N., Le Ber, F., 2019. Mining the sequential patterns of water quality preceding biological status of waterbodies. *Ecological Indicators* le 16 juin 2019, refusé le 22 août 2019 dans son format actuel, malgré « *son intérêt et son aspect innovant* » & proposition de le re-considéré dans un format plus allégé.

ARTICLES EN COLLABORATION

Dolques, X., Le Ber, F., Huchard, M., **Grac, C.**, 2016. Performance-friendly rule extraction in large water data-sets with AOC posets and relational concept analysis. *Int. J. Gen. Syst*. 45, 187–210. doi:10.1080/03081079.2015.1072927

Fabrègue, M., Braud, A., Bringay, S., **Grac, C.**, Le Ber, F., Levet, D., Teisseire, M., 2014. Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. *Ecol. Inform.* 24, 210–221. doi:10.1016/j.ecoinf.2014.09.003

CONFERENCE INTERNATIONALE

Grac, C., Dolques, X., Le Ber, F., Braud, A., Cernesson, F., Trémolières, M., Beisel, J-N., 2017b. A new data mining approach to understand the river ecological status: first large application of closed partially ordered patterns on French aquatic data. Oral presentation, *10° SEFS – Symposium for European Freshwater Sciences* – 2-7/07/2017, Olomouc, CZ.

CONFERENCE NATIONALE

Grac, C., Braud, A., Le Ber, F., 2018 – Extraction de motifs temporels caractéristiques des hydro-écorégions à partir de données de surveillance des rivières françaises. Communication orale, *Congrés AFL – Association Française de Limnologie* –, 22-23/11/2018, Strasbourg ;

CONGRES DES DOCTORANTS

POSTER

Grac, C., Berrahou, L., Bimonte, S., Boulil, K., Braud, K., Bringay, S., Cernesson, F., Dolques, X., Fabrègue, M., Fontan, B., Lalande, N., Le Ber, F., Levet, D., Molla, G., Niel, J., Teisseire, M., 2015. Une base de données pour l'évaluation et le suivi de la qualité hydrobiologique des cours d'eau. *Congrès des doctorants ED-413*, 30/11/2015, Strasbourg.

COMMUNICATION ORALE

Grac, C., Dolques, X., Le Ber, F., Braud, A ., Trémolières, M., Beisel., J-N., 2016. First large application of closed partially ordered patterns to data collected for the French national ecological assessment of rivers: toward a new approach to understand the state of aquatic ecosystems. *Congrès des doctorants ED-413*, 30/11/2016, Strasbourg.

SOMMAIRE

INDEX DES ANNEXES (disponibles en fin de documents dans une pagination propre)

INDEX DES ILLUSTRATIONS

INDEX DES TABLEAUX

INDEX DES TABLES

LISTES DES SIGLES & ACCRONYMES

ACID : Acidification, altération du SEQ-eau et des grilles DCE

AFB: Agence Française de Biodiversité

ANR : Agence Nationale de la Recherche

AZOT : Matières azotées hors nitrates, altération du SEQ-eau

BILO2 : Bilan oxygène, altération des grilles DCE

ENGEES : Ecole Nationale du Génie de l'Eau et de l'Environnement de Strasbourg

EPRV : Effets Prolifération Végétale, altération du SEQ-eau

DCE : Directive Cadre européenne sur l'Eau (2000)

HAP : Hydrocarbures aromatiques polycycliques, altération du SEQ-eau

HER: Hydro-écorégion

I2M2 : Indice Invertébrés Multi-Métrique

IBD: Indice Biologique Diatomique

IBGN: Indice Biologique Global Normalisé

IBMR : Indice Biologique Macrophytique en Rivière

ICUBE : Laboratoire des sciences de l'Ingénieur, de l'Informatique et de l'Imagerie (UMR 7357)

IPR: indice Poissons en Rivière

LIVE : Laboratoire Image Ville et Environnement (UMR 7362)

MINE : Paramètres de la minéralisation, altération du SEQ-eau

MOOX : Matières organiques et oxydables, altération du SEQ-eau

MPMI : Micropolluants minéraux, altération du SEQ-eau

MPOR : Micropolluants organiques hors pesticides, altération du SEQ-eau

NITR: Nitrates, altération du SEQ-eau

NUTRI : Nutriments, altération des grilles DCE

ONEMA : Office National de l'Eau et des Milieux Aquatiques

PAES : Particules en suspension, altération du SEQ-eau

PCB : Poly-chloro-biphényles, altération du SEQ-eau

PEST : Pesticides, altération du SEQ-eau

PHOS : Matières phosphorées, altération du SEQ-eau

POSPE : Polluants Spécifiques, altération des grilles DCE

PRESTOR: temporal PRESsure categories patterns extracTOR

RCS : Réseau de Contrôle de Surveillance

RM : Rhin Meuse

RMC : Rhône Méditerranée et Corse

SANDRE : Service d'Administration Nationale des Données et Référentiels sur l'Eau

SDP : Substances prioritaires et substances dangereuses prioritaires, altération des grilles DCE

SEQ-eau : Système d'Evaluation de la qualité de l'eau

TEMP : Température, altération du SEQ-eau et des grilles DCE

CHAPITRE I : INTRODUCTION

Nous sommes dans l'ère de la transition numérique : les données massives, désignées sous l'anglicisme « *big data* » associées à l'intelligence artificielle et ses algorithmes de fouille offrent des possibilités séduisantes et innombrables d'exploration de ces données dans de nombreux domaines tels que la médecine, la génétique, la sécurité. Mais aussi, le plus souvent, ces approches sont développées au service de notre société de consommation : plateforme d'achat aux prix évolutifs en fonction de la demande, publicités personnalisées à notre profil de consommateur, … L'actualité dénonce même des soupçons d'influences politiques personnalisées lors d'élections outre Europe. Parallèlement, nous vivons une 6ème putative extinction d'espèces et pour la première fois, nous et notre société de consommation en sommes responsables. Comment concilier ces deux facettes de notre monde actuel ? L'intelligence artificielle peut-elle être utilisée pour d'autres causes que le consumérisme, telles que la préservation des écosystèmes ? Cette puissance numérique peut-elle être mise au service de la compréhension des réactions des écosystèmes aux pressions qu'ils subissent dans le but d'identifier les leviers qui permettraient de les restaurer, de prédire leurs évolutions, **Ce travail est une contribution à cette tentative de conciliation entre écologie et informatique, appliquée aux écosystèmes rivières, en France.**

Après avoir posé le cadre actuel de la réglementation européenne de reconquête du bon état écologique des masses d'eau de rivière et de ses enjeux, cette introduction montre que les données disponibles sur les rivières sont à présent des données massives, analyse ce que peuvent apporter les méthodes de fouilles appliquées à ces données et enfin présente les objectifs de cette thèse.

Ce travail pluridisciplinaire associe hydrobiologie, écologie aquatique et informatique : **les objets et objectifs finaux appartiennent à l'hydroécologie et les méthodes employées sont informatiques**. Il est organisé en cinq chapitres

dont deux reprennent des articles soumis à des revues anglo-saxonnes. Dans les chapitres en français, nous avons tenu à éviter les anglicismes. Nous parlerons donc, notamment, de données massives pour désigner le « *big data* » et de fouille de données pour le « *data mining* ».

1 Le contexte de la réglementation européenne actuelle et ses enjeux: vers la reconquête du bon état écologique des rivières

Lors de la récente conférence de l'IPBES –Inter-governmental Science Policy Plateform on Biodiversity and Ecosystem Services- à Paris du 29 avril au 4 mai 2019, les scientifiques de 132 pays ont rappelé combien la biodiversité et les services qu'elle rend aux société humaines se détériorent (IPBES, 2019). Les écosystèmes d'eau douce n'échappent pas à cette érosion de la biodiversité malgré des politiques ambitieuses de maintien et de restauration de leur état que ce soit aux Etats Unis (US Clean Water Act, 1972 in Brogna, 2017) ou en Europe avec la Directive Cadre européenne (DCE) sur l'Eau 2000/60/CE (European Council, 2000).

L'objectif de ce chapitre est de définir la DCE, sa mise en œuvre, les bouleversements qu'elle a entrainés et les questions que son application pose encore.

Le vocabulaire spécifique à la DCE sera souligné à sa première apparition dans ce sous-chapitre. Ce vocabulaire est utilisé notamment dans les trois guides successifs d'évaluation de l'état des masses d'eau édités par le ministère de l'écologie en 2009, 2012 et 2016 (MEEDDAT, 2009; MEEM, 2016, 2012).

1.1 La DCE (Directive Cadre européenne sur l'Eau)

En Europe, la DCE impose une obligation de résultats : l'atteinte du bon état écologique par la conservation ou la restauration des masses d'eau à court et moyen termes. Trois cycles d'évaluation de six ans chacun étaient prévus : 2015, déjà échu, 2021 et 2027. La limitation des substances prioritaires et dangereuses prioritaires est également un objectif de la DCE : les masses d'eau devront être en bon état chimique, défini par rapport à ces substances. Ainsi, bien que ce soit le bon état

écologique qui est plus le souvent évoqué comme objectif de la DCE, ainsi que sa non dégradation, le texte vise bien à préserver ou restaurer un bon état global défini comme le pire des deux états écologique et chimique pour chaque masse d'eau. La DCE s'applique, sur tout le territoire français, outre-mer inclus, aux eaux douces continentales courantes, stagnantes, aux eaux de transition et aux eaux littorales. Nous ne parlerons ici que des eaux continentales courantes : les rivières, qui représentent la grande majorité des masses d'eau en France. L'échéance de l'atteinte du bon état écologique pour toutes les masses d'eau était 2015 ou sous réserve de contraintes techniques et/ou financières explicitées 2021, ou 2017. En France, 44,8 % des masses d'eau rivières ont été évaluées en bon état écologique pour la fin du premier cycle de la DCE (Blard-Zakar et Michon, 2018). Notre pays n'a pas réussi à atteindre l'objectif qu'il s'était fixé de 60 % de ses masses d'eau continentales en bon état pour 2015 comme fixé par la loi de programmation relative à la mise en œuvre du Grenelle de l'Environnement (Loi n° 2009/967 du 3 août 2009).

La masse d'eau est l'unité de base de la DCE : c'est elle qui est évaluée, puis en fonction de son état écologique, bon ou non, préservée ou restaurée. Pour les rivières, la masse d'eau peut être une rivière et ses affluents pour un petit bassin-versant, ou le plus souvent un tronçon de rivière, pour les rivières de grande taille. Une masse d'eau doit être homogène en taille, en conditions climatiques et géologiques mais aussi en fonction des principales pressions identifiées sur la base de l'occupation de son bassin-versant, telles que les grandes cultures ou la traversée de villes. La délimitation des masses d'eau en France s'appuie sur la circulaire n° 2005/1 du 29 avril 2005, relative à la typologie nationale des eaux de surfaces (cours d'eau, plan d'eau, eau de transition et eaux côtières) en application de la DCE. Chaque pays européen a dû définir la typologie de ses masses d'eau. La France a choisi de découper son territoire en hydro-éco-régions (HER) homogènes en conditions climatiques et géomorphologiques (J.-G. Wasson et al., 2004). Cette méthode a été considérée comme efficace au niveau européen (Verdonschot et Nijboer, 2004) et retenue par la majorité des pays. Certaines de ces HER ont un cas général et des cas dit « exogènes » d'une HER voisine. Ainsi l'HER 18, Alsace, correspondant à la Plaine d'Alsace a des masses d'eau appartenant au cas général et des masses d'eau du cas « exogènes » de l'HER 4, les Vosges, qui ont

commencé leur cours dans les Vosges et le terminent dans la plaine. Les masses d'eau auront des propriétés hydrologiques et de minéralisation différentes du fait des différences géologiques –plaine à substrat calcaire et Vosges à substrats gréseux et granitique- et hydrologiques –précipitations plus abondantes par formation orographique sur les Vosges. La France a défini cinq tailles de cours d'eau : très petit, petit, moyen, grand et très grand en fonction de leur rang de Strahler (MEEDDAT, 2009; Strahler, 1957) . Ainsi chaque masse d'eau appartient à une HER donnée et est d'une taille donnée. La DCE introduit également la notion de masse d'eau fortement modifiée : une masse d'eau sur laquelle des aménagements humains importants (digues, barrages, …) existent et ne seront pas remis en cause du fait d'enjeux forts économiques ou de de protection. Sur ces masses d'eau, seul le bon potentiel écologique est visé. Dans les masses d'eau non considérées comme fortement modifiées, il peut être décidé d'effacer les aménagements humains, c'est-à-dire de les supprimer s'ils sont la cause de la non atteinte du bon état écologique.

Que ce soit sur la remise en question de certains aménagements humains ou l'objectif du bon état écologique pour toutes les masses d'eau en Europe, la DCE représente un changement très fort d'orientation politique, aux niveaux européen et national, et sa mise en œuvre concrète a posé et pose toujours de nombreuses questions et défis à relever.

1.2 Les difficultés de sa mise en œuvre

Les difficultés de mise en œuvre de la DCE concernent ses trois étapes : d'abord l'évaluation de l'état écologique, en second lieu, l'identification des mesures de restauration et la capacité de résilience de la masse d'eau considérée, enfin, l'évaluation des effets des restaurations.

Paramètre, indice, métrique, indicateur, classe, seuils, état et qualité sont des termes très utilisés notamment dans le contexte de la DCE, mais à la définition et aux limites parfois floues. Dans ce document, nous utiliserons le terme de paramètre(s) pour les éléments physico-chimiques et les substances prioritaires et dangereuses prioritaires ; le terme d'indice(s) pour les indices biologiques, qui

peuvent être composés de plusieurs métrique(s) -excepté pour d'autres indices, tels que l'indice de similarité de Jaccard, mais nous le préciserons dans ce cas- ; le terme d'indicateur(s) soit comme terme générique désignant un ensemble de paramètres, indices et métriques, soit comme un résultat d'agrégation intermédiaire, tel que le percentile 90 annuel d'un paramètre chimique ; le terme de classe(s), pour les classes d'état ou de qualité obtenues après discrétisation d'indicateurs sur la base des valeurs-seuils existantes. Nous utiliserons le terme seuil(s) pour désigner ces valeurs-seuils. Enfin nous utiliserons le terme d'état lorsqu'il est évalué avec des seuils définis dans la cadre de l'application de la DCE et celui de qualité lorsque l'évaluation est faite avec des seuils qui ont été définis avant l'application de la DCE.

1.2.1 Evaluer l'état écologique

Selon la DCE, « *l'état écologique correspond à la qualité de la structure et du fonctionnement des écosystèmes aquatiques* » (MEEDDAT, 2009). En théorie, il doit être établi sur la base d'écarts à une référence du bon état préalablement défini pour chaque type de masses d'eau, dans chaque HER. Son évaluation doit être basée en priorité sur les compartiments biologiques, en considérant au moins un groupe animal et un groupe végétal. Cette notion a donné lieu à de nombreuses discussions à la fois nationales et européennes pour sa mise en œuvre ainsi qu'à plusieurs projets de recherche européens ou nationaux. Au niveau européen, des règles communes d'agrégation des différents éléments d'évaluation de l'état ont été fixées, un effort important de normalisation de diverses méthodes a été initié et continu toujours et un processus d'inter-étalonnage des indicateurs a été mis en place. Malgré cela, toutes les questions techniques ne sont pas encore résolues, comme exposé ci-dessous.

1.2.2 Une méthode commune d'agrégation des différents éléments d'évaluation de l'état

L'objectif de la DCE est donc de préserver ou restaurer un état global défini comme au moins bon, ce niveau étant le pire des deux états écologique et chimique pour chaque masse d'eau.

L'état chimique fait consensus au niveau européen : il est établi selon les mêmes règles, d'abord sur la base de 41 puis de 45 substances prioritaires et

dangereuses prioritaires définies par deux textes : la directive 2008/105/CE du Parlement et du Conseil Européen modifiée par la directive 2013/39/UE de l'Union Européenne. Les valeurs-seuils à respecter sont communes et existent à la fois pour les moyennes annuelles (MA) des concentrations mesurées et la concentration maximale annuelle (CMA). L'état chimique comporte deux classes : bonne et mauvaise, représentée par deux couleurs : bleue et rouge. Une seule substance dont la MA ou la CMA est au-dessus des valeurs-seuils tolérés suffit à déclasser l'état chimique.

Il existe cinq classes de l'état écologique : très bon, bon, moyen, médiocre et mauvais. Ces classes, représentées par cinq couleurs –bleue, verte, jaune, orange et rouge-, s'inspirent du modèle français mis en place avec le SEQ-eau (Système d'Evaluation de la Qualité de l'Eau) (MEDD et AE, 2003). Les indicateurs à prendre en compte et leurs valeurs-seuil pour définir cet état sont propres à chaque pays mais doivent respecter des règles communes : être basés sur au moins un groupe faunistique et un groupe floristique, et pour chacun d'entre eux prendre en compte la diversité et l'abondance des peuplements considérés, être adaptés à la typologie des masse d'eau, permettre la mesure d'un écart à une référence établie par type de masses d'eau. Ainsi la France a choisi que l'état écologique d'une rivière serait évalué sur la base de :

1) son état biologique, lui-même établi sur la base de quatre indices biologiques fondés : sur les invertébrés, indice IBGN (Indice Biologique Global Normalisé, AFNOR, 2004) que remplacera à terme l'I2M2 (Indice Invertébrés Multimétriques, Mondy et al., 2012), les poissons : indice IPR (Indice Poisson en Rivière, AFNOR, 2004b), les diatomées : indice IBD (Indice Biologique Diatomique, AFNOR, 2007) et les macrophytes : indice IBMR (Indice Biologique Macrophytes en Rivière, AFNOR, 2003),

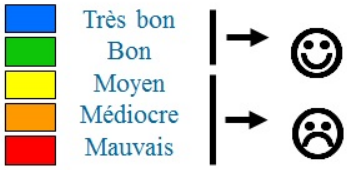
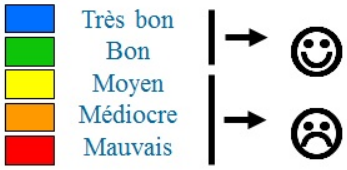
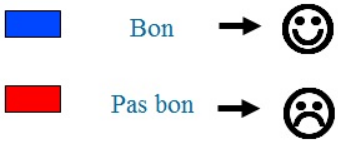
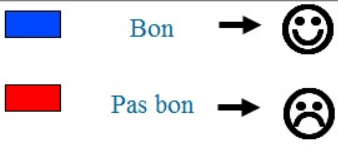
2) son état physico-chimique soutenant la biologie, composé d'éléments chimiques naturellement présents dans les eaux, tels que les nitrates, mais pouvant être impactés par les activités humaines,

3) les polluants spécifiques à l'état écologique, composé de quatre métaux lourds et de 4 à 16 molécules parmi 27 micropolluants synthétiques, tels que des pesticides, dont la liste varie selon l'Agence de l'Eau considérée.

Cette liste et les valeurs-seuils appliquées ont évolué depuis 2009. Ils sont à présent fixés dans le Guide Technique d'Evaluation de l'état écologique des masses d'eau (MEEM, 2016). Comme exigé par la DCE, ces règles d'évaluation ont également été fixées réglementairement par l'Arrêté 8 juillet 2018 (n° DEVO 1017167A) modifiant celui du 25 janvier 2010 relatif aux méthodes et critères d'évaluation de l'état écologique, de l'état chimique et du potentiel écologique des eaux de surface. Les règles d'agrégation annuelle des paramètres ayant plus d'une mesure par an, sont également propres à chaque pays. Ainsi pour les principaux paramètres physico-chimiques, hors substances de l'état chimique, la France a choisi le percentile 90 comme agrégation annuelle (MEEDDAT, 2009) alors que d'autres pays –Espagne, Royaume-Uni, Allemagne, Pologne, …- ont choisi la moyenne annuelle (Arle et al., 2011). Par contre la règle d'agrégation finale de tous les indicateurs est commune : un seul indicateur ne respectant pas sa valeur- seuil du bon état suffit à déclasser l'état écologique. En théorie, l'état hydromorphologique devrait également être pris en compte dans l'état écologique. Pour l'instant, ce n'est pas le cas en France, où un outil spécifique: CARHYCE (CARactérisation HYdromorphologique des Cours d'Eau) a été mis au point, mais pour lequel les règles d'évaluation de l'indicateur hydromorphologique global (IMG) obtenu ne sont pas encore finalisées (Tamisier et al., 2014).

Le Tableau 1 résume les éléments à prendre en compte pour l'évaluation de cet état global.

Tableau 1 : Eléments pris en compte pour l'évaluation de l'état global au titre de la DCE en France ; *l'état global étant le plus mauvais de tous (exceptés pour la physico-chimie qui ne peut déclasser l'état écologique au-delà de l'état moyen)*

Etat	Eléments de l'état	Définition des seuils	Grandeur à traiter	Nombre de classes
état écologique	**Hydromor-phologie**	nationale	Développement en cours	Non en core défini
	biologie	nationale	Valeur de 4 indices IBGN (à terme I2M2), IBD, IPR, IBMR	
	physico-chimique soutenant la biologie	nationale	Percentile 90 (ou 10) de 11 paramètres	
	polluants spécifiques de l'état écologique	Européen-ne	NQE_Ma* de 4 métaux lourds et de 4 à 16 molécules synthétiques suivant les Agences de l'Eau	
état chimique		Européen-ne	NQE_MA* & NQE_CMA* de 45 substances	

1.2.3 Des efforts d'homogénéisation de l'évaluation

Depuis la mise en place de la DCE, un important chantier de normalisation a été initié par le Comité Européen de Normalisation (CEN) pour partager les méthodes pouvant être mises en commun. Chaque instance de normalisation nationale participe à ces travaux et veille à harmoniser ses propres normes nationales aux normes européennes. Les normes ou projet de normes concernent les pratiques d'échantillonnage et de traitements des échantillons des groupes biologiques. Le plus souvent, les normes européennes restent très générales alors que les normes nationales sont à la fois plus précises techniquement et adaptées au contexte national.

Ainsi pour les échantillonnages en rivière, il existe à la fois des normes CEN et des normes de l'Association Française de NORmalisation (AFNOR). Par exemple, pour l'échantillonnage d'invertébrés en eau peu profonde : la norme EN 16150 (2012 Qualité de l'eau — Lignes directrices pour l'échantillonnage des macro-invertébrés benthiques en cours d'eau peu profonds au prorata des surfaces de recouvrement des habitats présents) et celle française NF T90-333 (AFNOR, 2016) ; ces documents ont notamment fait suite aux travaux du projet de recherche européen AQEM (Assessment system for the ecological Quality of streams and rivers throughout Europe using benthic Macroinvertebrates, 2000-2002) (Hering et al., 2004).

Des exercices d'inter-étalonnage des indices biologiques sont également organisés régulièrement afin que les évaluations soient comparables entre chaque pays et les efforts de reconquête du bon état écologique soient équitables. Les résultats du dernier exercice figurent dans la décision de la Commission n°2013/480/UE du 20 septembre 2013. Le projet de recherche européen STAR (Standardisation of River Classifications, 2001-2005) (Clarke & Hering, 2006) avait pour objectif de proposer une méthode générale d'utilisation des différents groupes d'organismes (macro-invertébrés, macrophytes, poissons et phytoplancton) pour l'évaluation de l'état des cours d'eau.

Malgré ces efforts à la fois de normalisation, de recherche et d'inter-étalonnage, les indices biologiques et leur mode de calcul sont propres à chaque pays. Théoriquement, c'est inévitable puisque ce sont des outils basés sur des peuplements d'êtres vivants ayant des aires de répartition biogéographique propres (Bonada et al., 2006; Borja et Dauer, 2008). Mais les raisons sont aussi pragmatiques : les pays ont utilisé leurs indices existants et dont ils tirent avantages à avoir des données anciennes, en les adaptant pour répondre aux exigences de la DCE. Aussi, les indices utilisés sont-ils très nombreux et ont été construits pour évaluer des degrés de dégradation, mais pas pour distinguer les pressions qui en sont la cause (Birk et al., 2012; Wasson, 2001).

1.2.4 *Identifier les pressions, évaluer la capacité de résilience des masses d'eau et l'effet des mesures de restauration*

Les gestionnaires attendent des outils permettant, pour un cours d'eau, d'identifier d'éventuels écarts à la référence et une capacité de résilience, de définir et de prioriser les actions de restauration et finalement d'évaluer les changements dus aux actions réalisées. Cependant ni les indicateurs biologiques actuels ni le mode d'évaluation globale, qui peut se résumer à une superposition d'évaluations des compartiments hydromorphologique, physico-chimique et biologique, ne répondent à l'ensemble de ces attentes.

Le cadre de travail et de réflexion préconisé pour l'application de la DCE est un concept dont l'acronyme anglais est DPSIR. Les activités humaines produisent des pressions qui altèrent les compartiments abiotiques et par conséquent impactent les communautés biologiques, et donc l'état écologique écologique, conduisant à une réponse politique d'amélioration de l'état si nécessaire (Concept DPSIR - *driving forces, pressure, state, impact, response* - de Kristensen, 2004). Les pressions subies par une même masse d'eau peuvent être multiples, avoir des effets synergiques ou à l'opposé antagonistes (Piggott et al., 2015). De plus, ces pressions peuvent s'exercer à différentes échelles spatiales (Dahm et al., 2013; Lalande, 2013; Villeneuve, 2016) depuis l'échelle stationnelle telle que la dégradation des micro-habitats, jusqu'à l'échelle du bassin-versant sous l'effet des pollutions diffuses issues de l'agriculte par exemple. Feld et al (2016b) proposent d'introduire la définition de "stresseurs" (e.g. concentration en nitrates et phosphates) plus spécifiques que celle de pression (e.g. pollution diffuse). Le stresseur a un lien direct cause-effet selon cette définition.

Or, distinguer et identifier les différentes pressions pesant sur une masse d'eau est la condition nécessaire pour choisir les éventuelles mesures de restauration à lui appliquer pour atteindre son bon état écologique, tout comme être capable d'évaluer sa capacité de résilience (Feld et al., 2016b; Reyjol et al., 2014).

Mais, comme montré précédemment, les indices biologiques utilisés ont été construits pour évaluer des degrés de dégradation, et non pas pour distinguer les pressions qui en sont la cause, encore moins dans des contextes multi-pollués, qui sont de plus en plus fréquents (Reyjol et al., 2014). Plusieurs travaux proposent de

compléter les résultats d'un indice biologique avec ceux des traits de vie du peuplement biologique échantillonné pour affiner le diagnostic des pressions, en s'appuyant notamment sur les travaux de Statzner et Bêche (2010). Ainsi Beketov et al. (2009) ont proposé l'indice SPEAR (Species At Risk) basé sur les invertébrés et leurs traits pour déceler en théorie des pressions aux micropolluants avec en pratique l'indice SPEAR pesticides spécifique aux pressions par les pesticides. Ce principe a été étendu plus largement aux conditions multi-polluées (Rasmussen et al., 2013). Mondy et Usseglio-Polatera (2013) avec les invertébrés, puis Larras et al. (2017) avec les diatomées proposent un outil de diagnostic pour 16 pressions potentielles, dont 10 physico-chimiques et six hydromorphologiques. En France, des travaux sont en cours sur le compartiment poissons (com.perso. P. Usseglio-Polatera, 12/09/2019).

Combiner les résultats obtenus sur différents compartiments biologiques pour distinguer les pressions semble une évidence, dès lors que l'évaluation selon la DCE généralise leur utilisation simultanée. Dès 2001, Lafont proposait le système d'ambiance écologique (EASY concept) pour distinguer les pressions mais aussi pour évaluer le potentiel de résilience de l'écosystème : les diatomées rendant compte de la qualité générale récente du compartiment eau, les macrophytes du niveau trophique de l'eau, les oligochètes de la qualité récente et ancienne vis-à-vis des pollutions organiques, ou toxiques du compartiment sédiment, les invertébrés et les poissons à la fois de la qualité du compartiment eau et des habitats, micro-habitats pour les invertébrés, macro-habitats pour les poissons. Il attribue une résilience possible des sédiments si la qualité moyenne de l'indice oligochètes n'est pas dépassée ; une résilience possible de l'ensemble du système si la qualité de l'indice invertébrés n'est pas dépassé. Il a appliqué ce concept aux indices IBGN, IBD, IBMR, IPR et IOBS (AFNOR, 2002) mais propose de l'étendre à d'autres indices basés sur ces mêmes groupes, dans d'autres pays. Ce système a été appliqué par Lafont et al. (2001) et Grac et al. (2006). D'autres auteurs ont comparé les réponses de plusieurs compartiments biologiques, excepté celui des oligochètes : s'ils s'accordent sur des réponses différentes des groupes biologiques aux différentes pressions, ils ne proposent pas de solution pour évaluer la capacité de résilience des écosystèmes étudiés. Les diatomées et les invertébrés sont donnés les plus sensibles à la qualité de l'eau (Marzin et al., 2012) et les poissons aux pressions

hydromorphologiques (Marzin et al., 2012; Villeneuve et al., 2015). Mais les conclusions peuvent varier en fonction des types de cours d'eau considérés : Hering et al., (2006, projet européen STAR -*Standardisation of River Classifications*, 2001-2005), puis Johnson et Hering (2009) démontrent qu'en montagne, ce sont les invertébrés et poissons qui répondent le mieux à la pression provoquant des excès de nutriments et à celles dégradant l'habitat, alors qu'en plaine, ce sont les diatomées et les macrophytes. Pour Villeneuve et al. (2015), quelle que soit l'échelle spatiale considérée -bassin-versant, tronçon ou station- les diatomées, invertébrés et poissons répondent bien aux pressions provoquant des excès de matières organiques et/ou nutriments, les poissons aux pressions dégradant l'hydromorphologie.

Enfin la dimension temporelle est évoquée pour le temps de réponse biologique, assimilé ici au temps d'intégration d'une pression par un compartiment biologique, qui est fonction des cycles de vie de chacun de ces compartiments. Les diatomées répondent en quelques semaines, les invertébrés en quelques semaines et jusqu'à plusieurs années, les poissons et les macrophytes sur des temps longs de quelques années, mais leurs temps de réponse en montagne peuvent être beaucoup plus courts (Johnson et al., 2006).

1.3 Synthèse et verrous

Ainsi, l'Union Européenne s'est donné les moyens d'appliquer la DCE en donnant un cadre commun et ambitieux pour préserver et restaurer le bon état des masses d'eau rivières, en instaurant des règles d'agrégation communes, en normalisant au niveau européen les grands principes des échantillonnages, par un exercice d'inter-étalonnage entre pays. Plusieurs projets de recherche depuis le lancement de la DCE ont également permis des avancées dans la typologie des masses d'eau, l'harmonisation des méthodes d'échantillonnage, l'introduction de nouveaux indicateurs répondant aux critères DCE, la proposition d'outils de diagnostic des pressions empêchant l'atteinte du bon état.

Malgré cela, le diagnostic des pressions via des modèles pressions-impacts, performants même en condition multi-polluée peut être amélioré. Evaluer la capacité

de résilience d'une masse d'eau est également un enjeu fort pour prédire le succès des restaurations. Explorer les réactions des masses d'eau sur de longues échelles de temps n'a pas encore été fait et pourrait apporter des éléments de réponses. Enfin les indices biologiques actuels peuvent être complétés et enrichis par l'utilisation des traits biologiques et écologiques, ou traits de vie, des taxons utilisés Travailler à l'échelle du multi-compartiment en combinant les indices biologiques est également à approfondir.

2 Les données disponibles sur les rivières : des données massives ?

2.1 Définitions des données massives

Si le terme « *Big data* » est largement utilisé, en trouver une définition bien établie est difficile. Nous retiendrons la suivante : « *le big data » décrit des ensemble de données volumineuses et/ou complexes qui ne peuvent être gérées et explorées qu'avec des techniques avancées »,* parmi celles recensées par Gandomi et Haider (2015) .

L'intérêt pour les données massives est récent, et est apparu avec le développement de la toile mondiale (*World Wide Web*) dans les années 1990 et celui des réseaux sociaux début 2000. Il est partagé à la fois par les institutions, les organismes de recherche, les entreprises qui y voient des opportunités d'avancées commerciales, sociales et d'innovation, non sans poser des questions éthiques que nous ne développerons pas ici. Mais la notion de volume n'est pas bornée, même si les volumes extrêmes atteignent à présent les Yotabytes (20^{24} bytes) et ne suffit pas à définir ces données massives. Depuis 2001, (Laney, 2001), la définition des données massives s'appuie sur trois dimensions: 1) leur Volume, 2) leur Vélocité et 3) leur Variété. La vélocité désigne des données qui arrivent en flux quasi-continu ; la variété désigne tant la diversité des sources de ces données que la diversité de leur format. A ces trois « V », peuvent s'en ajouter plusieurs autres. Nous n'en retiendrons ici que quatre autres : 4) leur Véracité : quelle confiance peut-on leur accorder ? 5) leur Variabilité dans le temps ; 6) leur Volatilité : lorsqu'elles arrivent en flux continus, comment les stocker et/ou les traiter efficacement ? 7) leur Vulnérabilité : comment garantir leur transfert de manière sûre, éventuellement confidentielle ? (conférence pédagogique de Gançarski P., 2018, https://www.ac-strasbourg.fr/ consulté le 17 mai 2019; Bouzeghoub et Mosseri, 2017).

2.2 Les données « eau » en France : des données massives ?

La surveillance des masses d'eau rendue obligatoire par l'application de la DCE génère au niveau européen de larges volumes de données hétérogènes,

d'origines multiples et de différentes échelles temporelles et spatiales (Hering et al., 2010).

En France, le premier réseau de surveillance des rivières a été mis en place à partir de 1970. Les paramètres surveillés ont évolué au fil des années, pour se stabiliser une première fois en 1994, à la création du Réseau National de Bassin –RNB- puis en 2005 et 2007, au moment de la mise en place des réseaux de référence pérenne –RRP- et de Contrôle et de Surveillance –RCS- en application de la DCE (Bouleau, 2007). Il existe plusieurs autres réseaux que nous ne détaillerons pas tous ici. Citons à titre d'exemple le réseau d'intérêt départemental (RID67) que le Département du Bas-Rhin a maintenu jusqu'en 2017 et sur lequel il suivait l'ensemble des éléments suivis sur le RNB, puis sur le RCS.

Les données relatives aux rivières peuvent être réparties en cinq types :

- les données milieux liées à la qualité de l'eau et des milieux aquatiques, acquises sur les réseaux de surveillance : paramètres physiques, physico-chimiques, ou indices biologiques ; nous désignerons ces données comme les données milieux dans le reste du document ;
- les données relatives aux stations de mesure : localisation, réseau d'appartenance, environnement… ;
- les données relatives au réseau hydrographique, ses caractéristiques physiques et les espaces qui lui sont associés : bassin versant défini à partir d'un exutoire, masse d'eau… ;
- les données relatives aux activités humaines qui se traduisent, par exemple, par des pressions sur le milieu : rejets ponctuels issus des stations d'épuration, pressions diffuses estimées à partir de données d'occupation et d'utilisation du sol ;
- les données relatives aux variables de forçage ou de contexte telles que les données climatiques, ou les données rendant compte de l'homogénéité hydroécologique, telles que les HER.

Les données milieux sont les plus complexes : elles peuvent être abiotiques et correspondre aux caractéristiques physiques ou physico-chimiques d'une station de rivière ou biotiques et s'appuyer sur l'échantillonnage d'une partie de la biocénose de la station considérée. Nous distinguerons les données brutes acquises sur le terrain, les données agrégées annuellement lorsqu'il existe plusieurs mesures par an, les données discrétisées en classe de qualité ou d'état. Les données physiques concernent l'hydromorphologie du cours d'eau : l'état des berges, du lit mineur et du lit majeur, l'existence de continuités longitudinales, latérales et verticales, d'annexes, les conditions hydrauliques (vitesse, géométrie du cours d'eau) et hydrologiques dont les débits.

Ces données sont de différentes natures : elles peuvent être quantitatives (nombre), semi quantitatives (indices, métriques, ou classes) ou qualitatives (description, texte). Elles peuvent être issues de mesures et/ou mobiliser de l'expertise. Elles sont spatialisées et peuvent être classées selon leur topologie: topologie ponctuelle pour les mesures de qualité d'eau par exemple, topologie linéaire pour le réseau hydrographique, topologie de surface pour les données d'occupation du sol, et les différents zonages. Ainsi une donnée correspondant à une concentration de nitrates faite sur une station de rivière sera spatialisée selon une topologie ponctuelle : l'eau a été prélevée en un point donné de la rivière. Par contre, une note d'indice biologique tel que l'IBMR faite sur un tronçon de rivière sera spatialisée selon une topologie linéaire. Enfin ces données ont également une dimension temporelle : elles peuvent représenter une situation instantanée, ou être valides pour une durée donnée, lorsqu'elles correspondent à des données agrégées. A chaque donnée, il convient donc d'y associer sa précision temporelle, spatiale et sémantique. La précision s'entend sur ses différentes dimensions : granularité, périodicité des mesures, étendue (emprise spatiale, durée des chroniques…).

Tableau 2 : Les éléments de qualité utilisés par la France avant l'application de la DCE : le SEQ –Système d'Evaluation de la Qualité

Eléments de qualité	Description	Outils pré-DCE	Mode d'agrégation et classes de qualité
Hydromor-phologie	Débit, caractéristiques physiques de la rivière	Projet d'un SEQ-physique non abouti	
physico-chimique	Macropolluants (ordre de grandeur mg/L) Paramètres d'origine naturelle pouvant être impactés par les activités humaines (ex. matières organiques, nitrates, …)	SEQ-eau	1 qualité physico-chimique 1) Agrégation annuelle de chaque paramètre: percentile 90 (ou 10 pour l'oxygène, percentile 90 et 10 pour le pH), application des seuils 2) Regroupement des paramètres en altérations et attribution de la qualité la plus déclassante 3) Qualité physico-chimique globale : celle de l'altération la plu déclassante Très bon Bon Moyen Mauvais Très Mauvais
	Micropolluants (ordre de grandeur µg/L) Métaux lourds (plomb, arsenic, …) d'originie naturelle ou anthropiques et molécules synthétiques issues des activités humaines (ex. pesticides, médicaments, …)		
biologie	5indices IBGN, IOBS, IBD, IPR, IBMR	SEQ-biologie	Une qualité par indice, pas d'agrégation finale Très bon Bon Moyen Mauvais Très Mauvais

Concernant les données physiques sur les milieux, avant la mise en œuvre de la DCE, la France avait développé un Système d'Evaluation de la Qualité, appelé SEQ, qui devait en théorie se décliner en 3 entités : le SEQ-physique pour l'évaluation de la qualité hydromorphologique, le SEQ-eau pour l'évaluation de la qualité physico-chimique de l'eau et le SEQ-biologique pour l'évaluation biologique. Avec le SEQ, la France a initié le principe de cinq classes de qualité allant de très bonne à très mauvaise, symbolisée par cinq couleurs allant du bleu au rouge ; idée reprise par l'Europe pour la DCE. Finalement le SEQ-physique n'a jamais abouti, le SEQ-biologique s'est limité à proposer des seuils nationaux pour chaque indice biologique dans chaque norme correspondante. Le SEQ-eau (MEDD et AE, 2003) est le seul qui a abouti. La qualité finale attribuée est celle du paramètre le plus déclassant. A la différence de la DCE, qui distingue « chimie » et « physico-chimie générale », le SEQ-eau regroupe tous les paramètres physico-chimiques, qu'ils

soient d'origine naturelle pouvant être impactés par les activités humaines (ex. matières organiques, nitrates, métaux lourds) ou d'origine artificielle (ex. pesticides, médicaments). Ces éléments sont résumés dans le Tableau 2 et le Tableau 4, qui listent respectivement les 14 altérations physico-chimiques prises en compte dans le SEQ-eau et les 6 prises en compte dans l'état physico-chimique soutenant la biologie et l'état chimique de la DCE.

Tableau 3 : Liste des 14 altérations physico-chimiques du SEQ-eau (version 2003) et de leur acronyme (nombre total de paramètres 181)

Altérations	Nb paramètres
MOOX (Matières organiques et oxydables)	7
AZOT (Matières azotées hors nitrates)	3
NITR (Nitrates)	1
PHOS (Matières phosphorées)	2
EPRV (Effets Prolifération Végétale)	2
PAES (Particules en suspension)	2
TEMP (Température)	1
ACID (Acidification)	2
MINE (Paramètres de la minéralisation)	8
MPMI (Micropolluants minéraux)	19
PEST (Pesticides)	68
HAP (Hydrocarbures aromatiques polycycliques)	15
PCB (Poly-chloro-biphényles)	8
MPOR (Micropolluants organiques hors PEST)	45

Tableau 4 : Liste des 6 altérations des états physico-chimiques soutenant la biologie et chimiques de la DCE (nombre total de paramètres 49)

Etat	Altérations	Nb paramètres
Physico-chimique soutenant la biologie (inclus dans l'état écologique)	TEMP (Température)	1
	ACID (Acidification)	1
	BILO2 (Bilan oxygène)	4
	NUTRI (Nutriments)	5
	POSPE (Polluants Spécifiques)	9
Chimique	SDP (Substances prioritaires et substances dangereuses prioritaires)	38

La France s'est dotée depuis 1992 (Bouleau, 2007) d'un référentiel national pour les données sur l'Eau, qui a évolué avec l'application de la DCE : le SANDRE - Service d'Administration Nationale des Données et Référentiels sur l'Eau (http://sandre.eaufrance.fr). Ce référentiel décrit les données et leurs méta-données associées, et propose des tables[1] déjà complètes pour certaines d'entre elles telles que la table des paramètres physico-chimiques, celle des taxons utilisés pour les cinq indices biologiques IBGN, I2M2, IBD, IBMR et IPR. Malgré ce référentiel commun, la qualité des données peut être encore hétérogène : erreur de référencement des coordonnées spatiales, incohérence des unités, … L'évolution des codifications dans le temps, principalement des taxons, pose également un problème de stabilité des données.

En théorie si toutes ces données sont publiques et accessibles, il faut le plus souvent accéder à plusieurs bases de données pour les télécharger. Nous ne détaillerons pas ici les serveurs permettant d'accéder aux autres données que celles des milieux : ils seront abordés dans le chapitre 2 lors de la présentation complète du premier jeu de données utilisés. Pour les données de milieux, les données physico-chimiques brutes et agrégées sont bancarisées par chaque Agence de l'Eau qui propose des solutions de téléchargement qui peuvent légèrement varier. Leur accès peut se faire via le portail national Eau France (https://www.eaufrance.fr/) qui renvoie sur les différentes agences. L'AFB a mis à disposition le serveur Naïades (http://naiades.eaufrance.fr/) qui permet de télécharger rapidement toutes les données brutes physico-chimiques. Concrètement, ce serveur récupère automatiquement les bases de données des Agences de l'Eau. A terme, ce service devrait permettre l'accès aux données biologiques, dont la bancarisation se limite encore aux résultats des indices biologiques et à leur transformation en classes d'état dans les bases de chaque Agence de l'Eau. La bancarisation des listes faunistiques et floristiques pour les milieux aquatiques est un problème non encore résolu en France. L'accès à ce type de données n'est possible qu'en consultant les services déconcentrés de l'Etat en charge de leur acquisition qui peuvent être suivant le groupe considéré : l'Agence de l'Eau, la DREAL (Direction Régionale de

[1] Table ici est emprunté au langage utilisé dans les bases de données ; une base de donnée est composé de différentes tables, chacune contenant un ensemble homogène de données, mises en lignes, et de leurs caractéristiques mises en colonnes.

l'Environnement, de l'Aménagement et du Logement) ou la délégation régionale de l'AFB. Leur format de stockage est très variable, mais ressemble le plus souvent à des fichiers excel ou équivalents. Nous avons réalisé ce travail de collecte et de bancarisation de données pour le premier jeu de données utilisés pour ce travail. Dans le cadre de l'application de la DCE, l'Etat français s'est appuyé sur de nombreux travaux de recherche notamment mené par l'IRSTEA (Institut National de Recherche et Technologies pour l'Environnement et l'Agriculture) qui a travaillé à bancariser une partie de ces données et dont dispose à présent l'AFB. C'est l'AFB qui nous a fourni ce jeu de données nationales que nous avons utilisé dans les chapitres 3, 4 et 5.

Quelques ordres de grandeur sont présentés ci-dessous (Tableau 5) sur les deux jeux de données utilisées dans la thèse. Ces jeux de données sont décrits en détail dans les chapitres 2 et 3. Le premier jeu correspond aux données disponibles sur l'ensemble des réseaux de mesures de l'Est de la France, sur les territoires des deux Agences de l'Eau Rhône-Méditerranée-Corse et Rhin-Meuse, ce qui correspond à 30% du territoire métropolitain. La taille de la base de données créée est de 20 Giga octets. Le deuxième jeu correspond aux données de la France entière pour le seul réseau RCS. La taille de la base de données créée est de 7 Giga octets.

Tableau 5 : Volume des deux principaux jeux de données utilisés

Jeu de données	**Est de la France (30% du territoire ; tous réseaux de surveillance confondus)**	**France entière (réseau RCS)**
Période	2000-2010	2007-2013
Nombre de fournisseurs de données	16	8
Nombre d'enregistrements totaux	60.10^6	24.10^6
Nombre de stations	7 975	1 781
Nombre de résultats physico-chimiques	14.10^6	23.10^6
Nombre de tables de la Base de Données	80	94
Taille	20 Giga octets	7 Giga octets

2.3 Synthèse et verrous

Ainsi les données eau en France remplissent cinq des sept « V » retenus pour définir des données massives. Elles sont Volumineuses, Variées de par leur format et leur complexité, Variables à la fois dans le temps, mais aussi dans l'espace. Elles peuvent être Volatiles : c'est le cas de la codification SANDRE des taxons qui peut évoluer dans le temps, mais également sur les données agrégées en classes de qualité ou d'état suivant les guides des valeurs seuils utilisés (Grilles SEQ-eau, Grilles DCE 2009, 2012, 2016 ; (MEDD et AE, 2003; MEEDDAT, 2009; MEEM, 2016, 2012)). Enfin, leur Véracité, c'est-à-dire leur qualité et donc leur niveau de confiance est à définir.

Par contre, n'ayant pas travaillé sur des données moissonnées régulièrement sur des bases de données nationales, mais sur des données limitées à des périodes temporelles (2000-2010 ou 2007-2013), nos données n'ont pas de caractère volatile, mis à part sur les aspects évoqués ci-dessus. Par ailleurs, comme il s'agit de données publiques, nous n'avons pas eu à gérer leur vulnérabilité.

Malgré leur caractère public, leur accès, leur format très hétérogène et leur qualité restent des enjeux forts. Enfin, quelle que soit l'effort de surveillance, les réseaux restent limités et la question de la qualité des bassins versants non surveillés, et donc sans données, se pose.

3 Apports des méthodes de fouille de données au diagnostic du bon état écologique

Ce paragraphe a pour objectif de définir la fouille de données, situer les méthodes de fouille que nous allons utiliser parmi les grands types qui existent et poser les verrous de ces méthodes quand elles sont appliquées au domaine des rivières.

3.1 Définition et principaux types de la fouille de données

La fouille de données désignent des techniques permettant d'extraire des régularités dans des données préparées et d'évaluer ces régularités (Dunham, 2003; Fayyad et al., 1996). Ces régularités peuvent être des règles, des corrélations, des dépendances. Elles peuvent être utilisées dans un objectif de description des données, d'explication ou de prédiction.

Le terme de fouille de données est apparu dans les années 90 afin de désigner les techniques d'exploration de données devenues massives. Mais si, à l'époque, le terme est nouveau, il englobe à la fois de nouvelles techniques -telles que la recherche de motifs (Agrawal et Srikant, 1995)- et d'autres déjà existantes, qui avaient été développées sur des jeux de données plus petits mais qui pouvaient être adaptées aux données massives. Ces techniques appartiennent à plusieurs domaines dont la délimitation peut varier d'un auteur à l'autre. Nous retiendrons ici ceux représentés en Figure 1 et simplifiés d'après Dunham (2003). Ces domaines appartiennent aux statistiques –dont les analyses multi-variées- et à l'informatique. Pour ce dernier, il s'agit de l'exploration des bases de données par requêtes, des apprentissages automatiques –tels que les arbres de décisions, les réseaux de neurones, les algorithmes génétiques- et d'autres méthodes –telles que les recherches de motifs, les règles d'associations, les treillis de Galois. Ces domaines ont évolué de façon parallèle. Les apprentissages automatiques (*machine learning* en anglais) font eux-mêmes partie de l'Intelligence artificielle, qui désigne l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine (définition du Larousse : https://www.larousse.fr). Au sens strict, l'appellation fouille de données est parfois

limitée aux nouvelles méthodes. C'est ce que nous utiliserons dans les autres chapitres du document.

Par ailleurs, quelles que soient les techniques de fouilles utilisées, deux catégories existent : les méthodes supervisées et celles non supervisées. Les méthodes supervisées requièrent un jeu de données d'apprentissage incluant à la fois les variables de départ et les résultats attendus qui permet de construire un modèle spécifique adapté à l'objectif visé. Une fois le modèle construit, il est testé sur un deuxième jeu de données, qui sert à le valider. Après validation, le modèle peut être appliqué à d'autres jeux de données pour prédire des résultats attendus dans des conditions similaires. Les méthodes non supervisées ne requièrent pas de jeux de données d'apprentissage. Elles explorent l'ensemble des données disponibles à la recherche de règles non connues par avance.

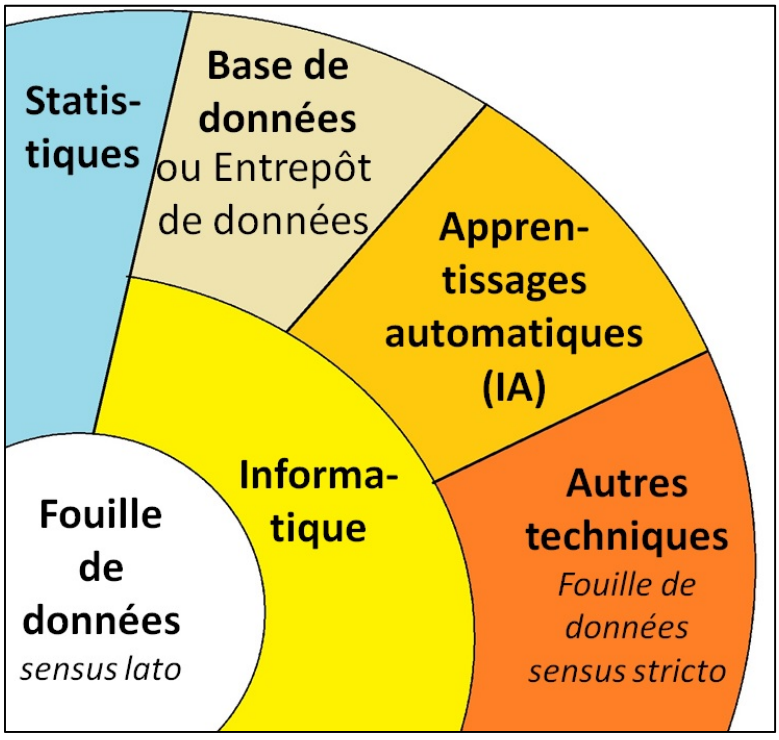


Figure 1 : Différents domaines de la fouille de données, au sens large et au sens strict (inspiré de Dunham, 2003)

3.2 La fouille de données : une des étapes de l'extraction de connaissances dans les bases de données (ECBD)

La fouille de données appartient plus largement à l'extraction de connaissances dans les bases de données (ECBD). D'après Fayyad et al. (1996), l'ECBD est un *« processus non trivial d'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données ».* Les étapes de l'ECBD sont 1) la sélection des données cibles incluses dans une base de données, 2) leur pré-traitement 3) leur transformation, 4) l'application de méthodes de fouilles, 5) l'interprétation et l'exploitation des régularités révélées par la fouille entre ces données 6) la validation de ces connaissances par l'utilisateur, expert du domaine (Figure 2).

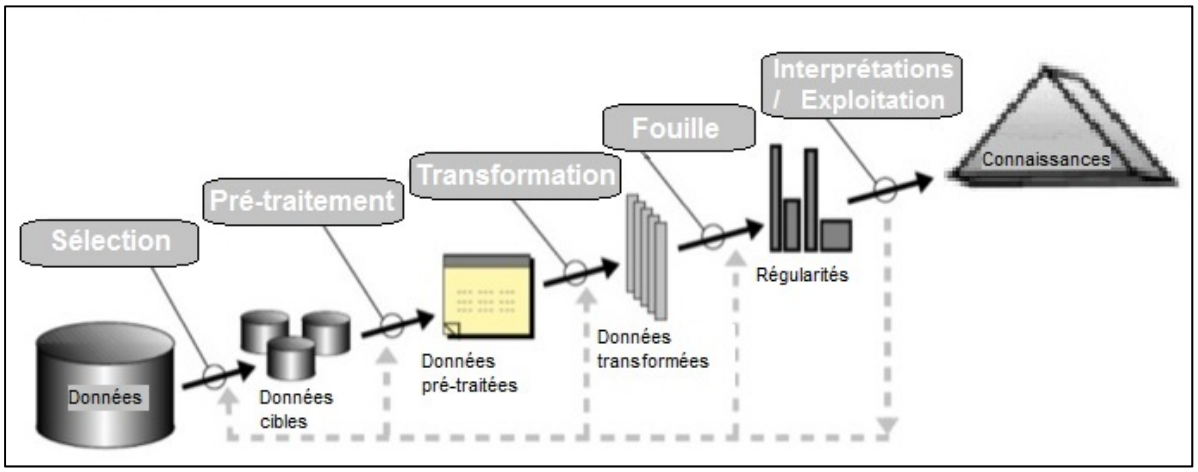


Figure 2 : Schématisation des étapes et des sous-étapes de l'extraction de la connaissance des bases de données (adapté de Fayyad et al., 1996)

Les types de données que nous allons utiliser ont été décrites au paragraphe précédent. Mis à part cette description des données, les deux étapes précédant la préparation des données et leurs fouilles n'ont pas été approfondies dans cette thèse. Notons toutefois que ces premières étapes sont essentielles : d'elles dépendent la fiabilité des résultats finaux. Ainsi, lors de la première étape de sélection des données, il est nécessaire d'évaluer leur qualité, en terme de complétude, cohérence, unicité, exactitude et fraicheur (Berti-Equille, 2004). Nous avons évalué la qualité dans les deux jeux des données utilisées (Berrahou et al., 2015) et les avons corrigées lorsque c'était nécessaires. La structuration des deux

jeux de données a également été menée avec rigueur : nous avons élaboré un modèle conceptuel de base de données au format SANDRE (Bimonte et al., 2015) en nous inspirant de travaux plus anciens (Grac et al., 2006). Pour le premier jeu de données, la base de données a été intégrée à un entrepôt de données (Boulil et al., 2014) incluant les grilles de qualité et d'état (MEDD et AE, 2003; MEEM, 2012). Cet entrepôt a été complété d'un outil de traitement analytique en ligne (équivalent anglo-américain : *On Line Analytical Process -OLAP*) permettant une analyse multidimensionnelle et multigranulaire spécifiquement adapté aux données de surveillance des rivières pour les données physico-chimiques et les données biologiques (Bimonte et al., 2015). Ainsi les résultats physico-chimiques peuvent être agrégés suivant les dimensions temporelle, spatiale, de validité - par exemple les analyses de nitrates, en 2007, en Alsace, supérieures à la limite de quantification de la méthode d'analyse choisie – et présentées dans différentes granularités –une classe d'état physico-chimique soutenant la biologie, une moyenne annuelle, …

3.3 Fouille de données et hydroécologie des rivières

La plupart des méthodes de fouille, au sens strict, sont capables d'analyser un grand volume de données, et peuvent être une bonne alternative aux méthodes traditionnelles de statistiques, notamment dans le domaine de l'écologie (Giraudel et Lek, 2001). Elles peuvent également être appliquées à de petits jeux de données (Bertaux, 2010). Ces méthodes peuvent produire des résultats lisibles, facilitant les interactions entre les informaticiens fouilleurs de données et les thématiciens d'un domaine (Džeroski et al., 1997).

Les hydroécologues ont eu recours à différentes méthodes pour explorer les données de surveillance des écosystèmes aquatiques, dont les rivières. En France, Verneaux (1973) a été le premier à utiliser les analyses multivariées avec l'Analyse Factorielle des Correspondances (AFC) élaborée par Benzécri (1973). Depuis, l'usage des méthodes statistiques multivariées s'est largement développé. En particulier en écologie, où les besoins croissant en bio-statistique, basée sur ces méthodes, ont donné naissance à des prologiciels (équivalent anglo-américain : *packages*) spécifiques dans l'environnement R (https://www.r-project.org/) tel que

ADE4 (Thioulouse et al., 1997). L'utilisation de méthodes d'apprentissage automatiques, fouille de données au sens strict, à ce type de données est plus récente. Walley et Džerosk (1996) ont testé trois de ces méthodes : les réseaux de neurones artificiels (RNA), l'inférence naïve Bayésienne et les arbres de régression et ont montré leur intérêt pour ce domaine d'application.

L'apprentissage par rétro-programmation, l'une des méthodes RNA, a été utilisé par Lek et al. (1995) pour estimer la consommation alimentaire par les populations de poissons et par Baran et al. (1996) pour prédire les caractéristiques des populations de truites selon l'habitat. En 1996, Walley et Džeroski ont testé les trois techniques d'apprentissage automatiques précitées sur des données de bio-surveillance aquatique. Ils ont démontré le potentiel de ces méthodes pour ce type de données. Les mêmes auteurs ont utilisé des arbres de régression sur des données relatives aux rivières britanniques et slovènes, pour prédire les paramètres physico-chimiques à partir de paramètres biologiques (Džeroski et al., 2000) ou à l'inverse, pour prédire le nombre de taxons à partir de paramètres physico-chimiques (Džeroski, 2001). D'autres méthodes ont été testées depuis ces travaux pionniers. D'heygere et al. (2003), ont utilisé des algorithmes génétiques pour prédire l'occurrence des macro-invertébrés benthiques dans les rivières en Flandre (Belgique). En 2006, ces auteurs ont élargi leur approche pour inclure d'autres outils : les arbres de classification et les réseaux de neurones artificiels (D'heygere et al., 2006), et plus récemment, les méthodes bayésiennes (Forio et al., 2016a; Landuyt et al., 2016b). Les réseaux de neurones artificiels ont été les premières méthodes d'exploration de données les plus utilisées. Lek et al. (1999), Imen et al., (2015), Markus et al. (2010), Xue et al. (2013) les ont utilisés pour prédire la concentration d'un seul paramètre chimique dans l'eau, les nitrates, dans les trois premières études citées et les matières en suspension dans la plus récente. Comte et al. (2010) ont utilisé les réseaux de neurones artificiels pour étudier la réponse des macro-invertébrés au stress environnemental dans le bassin de l'Escaut (Belgique). Les arbres de classification ou de régression arrivent en deuxième position (Xue et al.,2013; Nikoo et al., 2013). Poor et Ullman (2010) les utilisent pour prédire la concentration d'un seul paramètre chimique dans l'eau (nitrates ou matières en suspension). Dahm et al. (2013) et Feld et al. (2016a) ont utilisé des arbres de régression pour tester la réponse de plusieurs groupes biotiques (diatomées, macro-

invertébrés et poissons) aux contraintes environnementales (physiques et chimiques) dans les eaux courantes en se basant sur l'ensemble des données de surveillance allemandes et autrichiennes. Les arbres de régression ont été utilisés pour évaluer les pressions anthropiques sur les macro-invertébrés (Mondy et Usseglio-Polatera, 2013), les diatomées (Larras et al., 2017) ou sur les diatomées, les macro-invertébrés et les poissons ensemble (Villeneuve et al., 2015) sur les données de bio-surveillance françaises.

3.4 Synthèse et verrous

Les méthodes statistiques, incluses dans la fouille de données au sens large sont utilisées depuis de nombreuses années pour la recherche de liens entre pressions et réponses biologiques dans les écosystèmes aquatiques. Feld et al. (2016) ont proposé un « livre de cuisine » pour utiliser les méthodes statistiques adaptées aux données massives, comprenant une orientation pour l'analyse des données et l'interprétation des résultats, afin d'exploiter les données de bio-surveillance. L'application aux questions d'hydroécologie des autres méthodes de fouilles issues du domaine informatique, notamment celles non supervisées, est plus récente. Elles offrent un potentiel prometteur, mais leurs développements et adaptations aux questions liées à l'environnement et aux écosystèmes aquatiques en particulier nécessitent une collaboration étroite entre hydro-écologues et fouilleurs de données. Quelles sont les attentes des hydro-écologues? Quelles sont les possibilités et les limites des outils de fouilles de données ? Gilbert et al. (2018) ont proposé une évaluation de plusieurs méthodes de fouille de données, décrivant leurs principales caractéristiques et leur potentiel pour résoudre les questions environnementales. L'équivalent reste à faire dans le domaine du diagnostic de l'état des masses d'eau.

4 Les objectifs de la thèse

L'application de la DCE en Europe a renforcé la nécessité de diagnostiquer les pressions pesant sur les rivières, via des modèles pressions-impacts, qui puissent être performants même en condition multi-polluée. Evaluer la capacité de résilience d'une masse d'eau est également un enjeu fort pour prédire le succès des restaurations.

Nous avons décrit les données eau sur les rivières en France et montré que ce sont des données massives. D'autres auteurs avant nous ont expérimenté l'application des méthodes de fouilles, au sens large, à ces données : méthodes statistiques, réseaux de neurones. Par contre, l'application des méthodes de fouille au sens strict, non-supervisées reste limitée mais semble prometteuse. Par ailleurs, ce type de méthodes peut aussi s'appliquer à de petits jeux de données aux relations complexes, telles que des données de milieux sur les paramètres physico-chimiques, les listes de taxons associées à des données sur les traits biologiques et écologiques, ou traits de vie, de ces taxons.

Ce travail a pour objectif de répondre aux questions suivantes.

- Des méthodes de fouilles de données, principalement non supervisées, peuvent-elles permettre de diagnostiquer les pressions subies par les rivières et établir les liens avec leur état écologique ? La réponse à cette question nécessitera une collaboration étroite entre hydro-écologues et fouilleurs de données. Nous préciserons les étapes nécessaires à cette collaboration ainsi que les données indispensables qui à la fois respecteront les contraintes des méthodes et permettront l'obtention de résultats adaptés aux données, concrets, utilisables et transposables. Trois questions hydroécologiques parmi huit initiales ont été retenues. Une méthode de fouilles différente a été appliquée à chacune. Notre terrain d'étude est l'Est de la France (données de surveillance des rivières des deux agences de l'Eau Rhône-Méditerranée-Corse et Rhin-Meuse) (chapitre 2).

- A l'échelle nationale, les successions de pressions temporelles sur plusieurs mois à années permettent-elles d'expliquer les réponses des différents indices biologiques à

un instant donné? Certaines de ces successions peuvent-elles être identifiées comme caractéristiques de réponses types et donc servir à prédire une réponse biologique? Cette question était une des trois questions précédentes sélectionnées. Nous avons étendu le terrain d'étude à l'ensemble des données de surveillance de la France métropolitaine. La méthode de fouille utilisée est l'extraction de motifs séquentiels partiellement ordonnés dans le temps. Nous proposons d'optimiser cette méthode pour obtenir la sélection des motifs pertinents dans le cas de motifs extraits pour des durées de 24 mois, pour deux indices biologiques : l'IBGN et l'IBMR (chapitre 3).

- La durée des successions temporelles influentes varie-t-elle d'un indice biologique à l'autre? Sur le même jeu de donnée national, nous analyserons pour cela les motifs extraits pour cinq durées différentes 3, 6 12, 18 et 24 mois, pour les cinq indices biologiques dont nous disposons : l'IBGN, l'I2M2, l'IBMR, l'IBD et l'IPR (chapitre 4).

- La réduction de l'échelle spatiale permet-elle d'allonger la durée considérée des successions de pressions ? comment affiner et automatiser la sélection des successions les plus pertinentes ? Pour cela, nous avons travaillé à l'échelle d'une HER, l'Alsace, et pour le seul indice biologique I2M2 (chapitre 5).

- A l'échelle spatiale d'une HER, existe-t-il des successions de pressions physico-chimiques caractéristiques d'un changement d'état biologique ? Pour cela, nous avons travaillé sur l'HER Alsace et cherché des motifs caractéristiques de changement d'état du seul indice biologique I2M2 (chapitre 5).

L'ensemble de ces investigations permettra de dresser un bilan du potentiel d'utilisation des méthodes de fouilles non supervisées pour apporter des éléments originaux sur le diagnostic qui sous-tend l'évaluation des masses d'eau.

CHAPITRE II : Quelles méthodes de fouille de données utiliser pour évaluer l'état écologique des rivières ?

1 Résumé élargi

Ce chapitre a été soumis au journal Enviromental Modelling and Software le 22 août 2019, sous le titre "Which data mining method to use for the evaluation of river ecological status – Feedback on a close collaboration between data scientists and hydro-scientists". Il s'agit de la troisième soumission. Il avait été soumis une 1ère fois à Sciences of Total Environnement le 25 mai 2017 et refusé le 22 juin 2017 car ne correspondant pas aux attentes de la revue. Nous l'avions repris, puis soumis une 2ème fois à Plos One le 6 octobre 2017. Plos One l'avait refusé le 2 août 2018, suite à la demande de révisions majeures par les relecteurs et « malgré son intérêt et l'important travail réalisé ».

Nous avons montré dans l'introduction que les données de surveillance de l'état des masses d'eau en Europe, suite à l'application de la DCE, sont à présent des données massives. Nous souhaitons tester ici si des méthodes de fouille de données, dont des non supervisées, peuvent permettre de diagnostiquer les pressions subies par les rivières et établir des liens avec leur état écologique. La complexité des questions liées à l'évaluation de l'état des hydrosytèmes a nécessité une collaboration étroite entre chercheurs informaticiens, spécialistes des données et de la fouille de données, et thématiciens : hydrologues et hydroécologues, à la fois chercheurs et professionnels de deux bureaux d'étude.

Notre démarche s'apparente à une extraction de connaissances des bases de données (ECBD), mais pour rendre la collaboration pluridisciplinaire fructueuse nous préconisons une démarche non pas linéaire comme proposé dans l'ECBD (figure 2, chapitre I) mais un processus itératif en cinq étapes (Figure 3):

1. la définition d'une question thématique, sa traduction en question formalisée par les informaticiens, la sélection des données correspondantes ainsi que la définition de résultats types attendus ;
2. le choix de la méthode de fouille appropriée à la question formalisée;
3. la préparation des données, par agrégation, et éventuellement par discrétisation ;
4. l'application de la méthode, la production des résultats et leur hiérarchisation pour sélectionner les plus pertinents grâce à des mesures d'intérêt appropriées; les méthodes de fouille peuvent produire de très nombreux résultats et cette étape de sélection est cruciale ;
5. la validation et l'interprétation par les thématiciens des résultats sélectionnés : les résultats attendus servant à valider la méthode et les mesures d'intérêt choisies, les résultats inattendus permettant une nouvelle extraction de connaissances grâce à la méthode de fouille choisie.

Le processus est itératif : chaque étape est reprise autant que nécessaire de manière à affiner la démarche globale. Ce fonctionnement nous a permis d'adopter un langage commun entre informaticiens, hydrologues et hydro-écologues, de comprendre et maitriser les attentes et les contraintes de chacun.

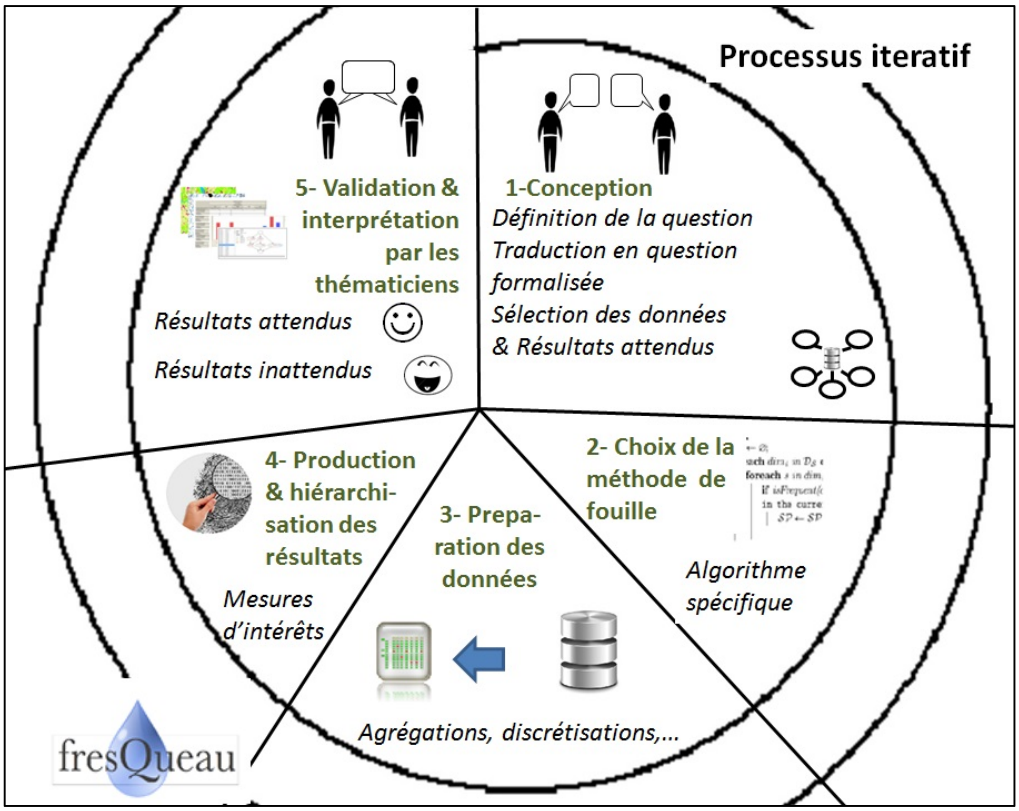


Figure 3 : Schéma du processus itératif proposé d'application de la fouille de données à des questions d'évaluation des rivières

Notre terrain d'étude est l'Est de la France : nous avons collecté et organisé en base de données, respectant le format SANDRE (op.cit.), les données issues de 21 sources différentes. Il s'agit de celles de l'ensemble des réseaux de surveillance physico-chimiques et biologiques des rivières des deux agences de l'Eau Rhône-Méditerranée-Corse et Rhin-Meuse, pour la période 2000-2010, ainsi que les données associées – localisation géographique des stations de mesures, réseaux hydrographiques, activités humaines, données de forçage telles que le climat ou les hydro-éco-régions. La plupart des données étaient publiques, mais pas toutes faciles d'accès et certaines étaient stockées sur des supports multiples tels que des fichiers de tableurs ou des extraits partiels de bases de données, …

Nous illustrons notre démarche en appliquant trois méthodes de fouille aux trois questions suivantes s'intéressant à trois dimensions des données dont nous disposons : leur temporalité, leurs liens relationnels et leur spatialité.

Question 1 : existe-t-il des relations temporelles entre les résultats physico-chimiques et les indices biologiques en un même lieu? La question formelle retenue par les informaticiens est la recherche de motifs dans des séquences de données temporelles. Elle a été appliquée à l'ensemble des données physico-chimiques et biologiques dont nous disposions. Les données physico-chimiques ont été agrégées en altérations du SEQ-eau (MEDD et AE, 2003) et discrétisées en classes de qualité. La méthode de fouille choisie est l'extraction de motifs fermés partiellement ordonnés. Il s'agit d'une méthode non supervisée. Nous l'avons spécifiquement adaptée pour répondre à la question (Fabrègue et al., 2014). Seule la durée de un an précédant un indice biologique dans un état donné a été testée. Nous obtenons des motifs, relativement didactiques, tels que la Figure 4. La plupart des altérations trouvées semblent en cohérence avec les indices correspondants, comme ici les particules en suspension et les matières organiques précédant un IBGN moyen. Nous constatons le plus souvent un décalage de classe entre les altérations physico-chimiques et les indices biologiques, les seconds étant en plus mauvais état que les précédents. Les mesures d'intérêt proposées pour filtrer les résultats sont la fréquence, la discriminance et la redondance de ces motifs, mais ils restent à affiner car les résultats à interpréter par les thématiciens demeurent encore très nombreux.

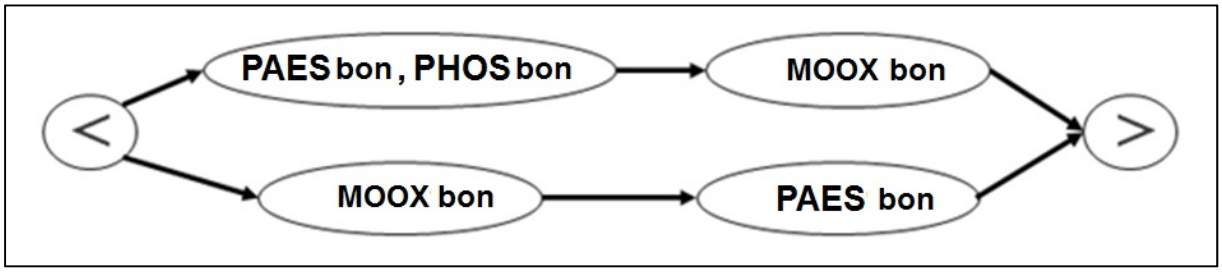


Figure 4 : motif fermé partiellement ordonné vérifié pour toutes les séquences temporelles de un an précédent un IBGN d'état moyen (altérations PAES : particules en suspension, MOOX : matières organiques et oxydables).

Question 2 : les variations des paramètres abiotiques d'une station de rivière peuvent-elles être reliées aux variations de fréquence d'expression de modalités de traits biologiques et écologiques des êtres vivants qui s'y trouvent ? La question formelle retenue par les informaticiens est la recherche de règles d'implication. La

méthode de fouille choisie est l'analyse relationnelle de concepts. Elle exploite les relations entre les paramètres abiotiques – hydromorphologiques et physico-chimiques – d'un site, les invertébrés qui y sont échantillonnés et leurs traits de biologiques et écologiques. Nous l'avons appliquée sur le seul territoire de l'Alsace. Il s'agit d'une méthode non supervisée. Nous l'avons spécifiquement adaptée pour répondre à la question (Dolques et al., 2016). Les données ont été discrétisées : celles abiotiques en classes de qualité, les abondances des invertébrés en quartiles définissant quatre classes de faible à très abondante, les fréquences d'expression des modalités des traits en trois niveaux d'affinités : faibles, moyennes et fortes. Les règles obtenues sont hiérarchisées en fonction des pourcentages de stations les vérifiant. Ces pourcentages restent faibles (maximum 15%) mais s'expliquent par le très grand nombre de possibilités (22 traits, ayant chacun de 2 à 9 modalités possibles, soit 119 modalités). La règle trouvée « *si une abondance élevée de taxons ayant une forte affinité pour la modalité potamon du trait écologique de distribution longitudinale est trouvée, alors les nitrites sont de qualité moyenne* » peut être considérée comme un résultat attendu validant la méthode sur notre jeu de données où les zones de potamon peuvent présenter de fortes concentration en nitrites (moyenne: 0,14 mg/L, maximum 1,54 mg/l). La règle suivante « si une *abondance élevée de taxons ayant une forte affinité pour la modalité lente du trait écologique préférence de courant est trouvée, alors l'état de l'hydrologie est médiocre »* a été interprétée comme un résultat inattendu mais correspondant, en Alsace, aux stations des anciens bras déconnectés du Rhin qui ont perdu leur dynamique hydrologique suite aux aménagements de fleuve. Cette méthode apparait donc comme adaptée à la question posée, mais nous n'avons pas encore résolu le problème de changement d'échelle pour l'appliquer à un territoire plus grand et un plus grand nombre de taxons – poissons, diatomées, macrophytes – sans avoir une explosion du nombre de règles obtenu. Sur seulement les données de l'Alsace, nous avions généré 1 428 règles et retenu 189 d'entre-elles.

Question 3 : quelles sont les relations entre les pressions liées à l'occupation du sol et les résultats de l'indice biologique IBGN ? La question formelle retenue par les informaticiens est un problème de modélisation spatiale. La méthode de fouille choisie est la corrélation spatiale de proches voisinages à laquelle nous avons dû intégrer la dimension orientée des réseaux hydrographiques (Lalande et al., 2014).

Elle appartient aux bio-statistiques. Elle a été appliquée aux données d'occupation du sol et aux résultats IBGN sur deux sous-bassins-versants de la Saône. Nous ne détaillerons pas plus, ici, cette méthode qui est hors champ des méthodes de fouille non supervisées et qui a été mise en œuvre à Montpellier.

L'extraction de connaissances dans les bases de données (ECBD) utilisant des méthodes de fouille, pourrait répondre aux besoins de stockage et d'analyses de ces données et aider à la décision les gestionnaires de ces masses d'eau. Cependant pour être efficace, cette approche doit être une collaboration pluridisciplinaire étroite entre les thématiciens et informaticiens spécialistes de la fouille. Notre travail peut servir de feuille de route pour mener à bien ce type de collaboration et pourrait être appliqué à d'autres domaines

2 Article

Which data mining method to use for the evaluation of river ecological status – Feedback on a close collaboration between data scientists and hydro-scientists

Authors: Corinne Grac[a,h], Flavie Cernesson[b], Xavier Dolques[c], Agnès Braud[c], Agnès Herrmann[e], Frédéric Labat[f], Maguelonne Teisseire[g], Michèle Trémolières[h], Florence Le Ber[c]

a ENGEES, F-67000 Strasbourg, France
b TETIS, AgroParisTech, CIRAD, CNRS, Irstea, Univ Montpellier, Montpellier, France
c ICube UMR 7357, Université de Strasbourg, ENGEES, CNRS, F-67400 Illkirch-Graffenstaden, France
e LHYGES UMR 7517, Université de Strasbourg, ENGEES, CNRS, F-67000 Strasbourg, France
f Aquabio firm, F-33750 Saint-Germain-du Puch
g TETIS, Irstea, AgroParisTech, CIRAD, CNRS, Univ Montpellier, Montpellier, Franceh LIVE UMR 7362, Université de Strasbourg, CNRS, F-67000 Strasbourg, France

Corresponding author: Corinne Grac corinne.grac@engees.unistra.fr

2.1. Highlights

1. Assessing the ecological status of European running waters generates "big data".
2. Data from one third of the French territory were modelled in a 20 Go database.
3. Hydro-scientists collaborated with data scientists in the project.
4. Iterative discussion was used to explore the data in an interdisciplinary approach.
5. The data mining methods used produced very promising results.

2.2. Abstract

The assessment of waterbodies relies on monitoring, generating large volume of data. The aim of the FresQueau project (2011-2015) was to develop and apply data mining tools to these data to understand the functioning hydro-ecosystems and to evaluate their ecological status. We were data scientists and hydro-scientists. We work in an iterative knowledge discovery process: question design, selection and preparation of the data, choice and adaptation of suitable data mining methods, analysis of the results. Results analysis allowed both to validate and to refine the methods, tuning the different steps of the process. We collected and organized public data originating from the monitoring of national waterbodies in the east of France. Three training cases are described, from the question definition to the results analysis, involving three aspects of the data: temporal, relational, and spatial. This article can be read as a roadmap for successful collaboration between data scientists and hydro-scientists.

2.3. Keywords

interdisciplinary approach, waterbodies' ecological status, database, discriminant temporal patterns, relational concept analysis, correlation tests.

2.4 Introduction

The European Water Framework Directive (WFD) (European Council, 2000) requires the achievement of a good (ecological and chemical) status for the conservation or restoration of waterbodies in the short (2021) and medium term (2027). According to the WFD, for running waters, waterbody "means a discrete and significant element of surface water such as […] a stream, river or canal, part of a stream, river or canal […]". The WFD defines ecological status as "an expression of the quality of the structure and functioning of inland aquatic ecosystem" (WFD, art. 2). Managers expect tools to (1) identify differences between the status of a given waterbody and its reference, as well as its potential of resilience, (2) define and prioritize restoration actions and (3) assess changes due to actions undertaken. Since the application of the WFD, the assessment of waterbodies has yearly generated large volumes of heterogeneous and multi-source data (Hering et al., 2010). European countries thus require tools to help organize and analyse these data, in order to feed decision-making systems for managers (Reyjol et al., 2014). The diversity, complexity and heterogeneity of the data, their temporal and spatial character, raise several issues that are specific to the management and analysis of environmental data. These issues are more in line with recent data mining approaches than with traditional data analysis approaches. Data scientists are needed who can organize and analyse these large volumes of data, known as "big data".

Methods for exploring data to monitor the status of aquatic ecosystems have been developed for many years. In France, Verneaux (1973) was the first to use multivariate analysis (factorial correspondence analysis, FCA) developed by Benzécri (1973). Since then, statistical methods have been widely used by hydro-scientists. Increasing needs in biostatistics led to the development of specific packages in R, as the ADE4 package (Thioulouse et al., 1997). The use of machine learning methods in hydroecology is more recent. Walley and Džeroski (1996) tested three machine learning methods on aquatic biomonitoring data: artificial neural networks (ANNs), naive Bayesian inference and regression trees and demonstrated the potential of these methods for this kind of data. ANNs were the first most widely used, e.g. to

estimate food consumption by fish populations (Lek et al, 1995) or to predict the characteristics of trout populations according to habitat. Lek et al. (1999), sed ANNs to predict the concentration of nitrate in natural water, while Comte et al. (2010) studied the response of macroinvertebrates to environmental stress. Classification or regression trees came in second: Poor and Ullman (2010) used them to predict the concentration of a single chemical parameter in natural water (nitrate or total dissolved solids). Dahm et al. (2013) and Feld et al. (2016a) used boosted regression trees to study the response of several biotic groups (diatoms, macroinvertebrates and fishes) to environmental (physical and chemical) stresses in running waters using German and Austrian monitoring datasets. Regression trees were used to assess anthropogenic pressures on macroinvertebrates (Mondy and Usseglio-Polatera, 2013), diatoms (Larras et al., 2017) or on diatoms, macroinvertebrates and fishes together (Villeneuve et al., 2015) using French national monitoring datasets.

Recently, surveys were made on which methods and which questions can match. Indeed, mutual understanding between scientists in hydro-ecosystems, hereafter 'hydro-scientists', and data scientists is indispensable to answer questions such as: what are the specificities of the data? What do the hydro-scientists expect? What are the possibilities and the limits of the chosen data mining tools? Gibert et al. (2018) proposed an assessment of several data mining methods, describing their main characteristics with respect to environmental questions. Feld et al. (2016b) proposed a 'cookbook', including a guidance for data analysis and result interpretation, to exploit biomonitoring data.

Besides, to go further than developing and applying various data mining methods on available data, data scientists have introduced the notion of "Knowledge Discovery in Databases" (KDD), defined as follows: "KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process" (Fayyad et al., 1996). The KDD process includes additional steps: data selection, data preprocessing (e.g. cleaning), data transformation (e.g. discretization), interpretation, and evaluation of data mining results. All steps are iterative, e.g. after the evaluation step, the selection or transformation steps can be reitered in order to improve or refine the results. Overall, KDD is a user-driven process, in opposition to blind application of data mining methods.

The aim of the 4 year-long ANR-11-MONU-14 FresQueau project (http://dataqual.engees.unistra.fr/fresqueau_presentation_gb) (2011-2015), was to prove the potential of data structuring and data mining methods for the analysis of all available data on French running waters. The FresQueau project had two specific aims: (1) to highlight links between different metrics which would make it possible to characterize the status of watercourses and (2) to link the sources of pressure on the environment to the physical-chemical and biological quality of running waters. Answering these questions required the use of several data sources related to water quality: hydrology, sampling sites location, etc., to the environment of the watercourses: land use for example. One of the greatest difficulties was the high heterogeneity (numerous types, formats, sizes, with potential temporal or topological dependencies) of these data. All the data used are public and cover one third of the French metropolitan territory including Corsica (the East of France). The FresQueau consortium included data scientists and hydro-scientists (hydrologists, hydro-ecologists –specialized on macroinvertebrates, macrophytes and biomonitoring of running waters, researchers as well as practitioners from two small firms). We used a methodological and iterative approach to understand each other, and to find out original methods that match environmental questions. As such, our work fits the KDD framework.

Here we present our working approach and illustrate its application with three training cases, in which three different data mining tools were applied to three topics of relevance to WFD: 1) Discriminant temporal patterns to link physical-chemistry and biology for the assessment of hydro-ecosystems, 2) Relational concept analysis to search for links between the life history traits of macroinvertebrates and the status of the hydro-ecosystem and 3) Multiscale spatial modelling to link land uses and the ecological status of hydro-ecosystems. We first describe the data we used and the modelling we performed to gather the data in a single database. In the discussion section, we 1) describe our approach limits due to the availability and quality of data, 2) focus on its innovative and performance aspects, and 3) highlight the role of interdisciplinarity to deal with issues raised by the conservation and restoration of waterbodies.

2.5 Material and methods

2.5.1. Data

The study area covered 161,100 km² in the east of France (Figure 5), representing 29.45% of the French metropolitan territory. The watersheds concerned are grouped in two major hydrographic areas: Rhine-Meuse (north-east; 33,000 km², 7,000 km of watercourses), and Rhone Mediterranean Corsica (south-east; 130 000 km², 152,000 km of watercourses).

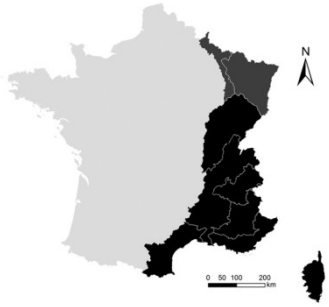


Figure 5: Location of the study area, covering eastern France (dark)

Our study period was 2000-2010. We collected five categories of water data from 21 different sources: mainly from administrations or public data banks and from some research projects. The five categories of data were: 1) river quality data, 2) sampling site data; 3) hydrographic network data, 4) human activities data, and 5) driving data (Figure 6). Below we describe each of the five categories and their main characteristics.

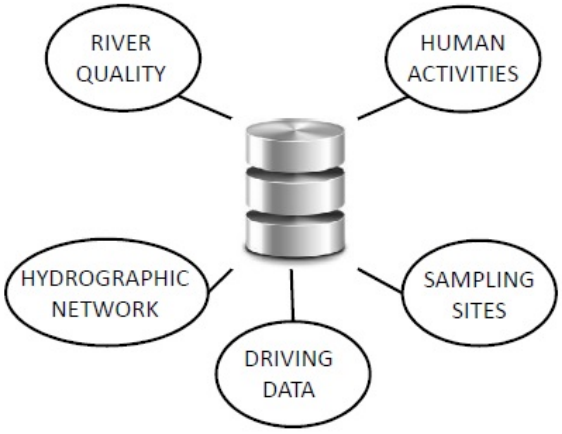


Figure 6: The five categories of FresQueau data

River quality data were divided into three sub-categories: physical (e.g. the dimensions and shape of the bed, characteristics of the substrate, condition of the banks), physical-chemical (chemical parameters of the water, e.g. pH, nitrate and phosphate, pesticides in the water or sediments…) and biological data (i.e. lists of fauna and flora taxa; biocenotic metrics, e.g. total abundance, diversity and biological indices) for four groups: macroinvertebrates, diatoms, macrophytes and fishes. The biological sampling, identification and calculation of the indices were standardized (i.e. IBGN, AFNOR (2004a) for invertebrates; IBD, AFNOR (2007) for diatoms; IBMR, AFNOR (2003) for macrophytes and IPR, AFNOR (2004b) for fishes). Associated with the taxa, we collected their life history traits (e.g. type of respiration, habitat preference), from the database built for a previous research project 'Indices' (Grac et al., 2006). These traits were defined by Usseglio-Polatera et al. (2000) for invertebrates, Van Dam et al. (1994) for diatoms and Bornette et al. (1994), Willby et al. (2000) for macrophytes.

Sampling site data gave the location of the sites where the river quality data were sampled and their main characteristics (e.g. denomination, objectives, biomonitoring network type, producer). The geometry of a sampling site was considered as a point, whereas it could be a point (for sampling of physical-chemical characteristics and diatoms) or a segment (for sampling of other biological groups and physical parameters). First, we had to convert all the geographical coordinates into the same system of projection (Lambert 93 system).

Hydrographic network data were geographical data: the different segments of running waters or of waterbodies, the size of watersheds, administrative regions, and complementary information on the hydrographic network.

Human activity data such as land use (i.e. natural, urban, type of agricultural land use), impediments to flow, location of discharges (e.g. discharges from waste water plants) to estimate anthropogenic pressures on running waters. These pressures may be intermittent or permanent (e.g. discharge of a waste water plant), chemical (e.g. organic matter from poorly treated waste water) or physical (e.g. a dam on a river which prevents the free circulation of fishes). These pressures may also be diffuse (e.g. leaching of agricultural inputs such as nitrate).

Driving data concern forcing or context variables such as climate (e.g. average atmospheric temperature, precipitation), flows, hydro-ecoregions or administrative information to characterize the environment of the running waters and sampling sites.

There was considerable variability in the sources, which often depend on dataset owners' objectives and may differ over time and in space, as well as in format. These data were highly heterogeneous: simple measures of a parameter (e.g. a pH value) versus a complex index using different metrics (e.g. the French macroinvertebrate index, IBGN, op. cit.) or based on expert knowledge. To this heterogeneity, are added the diversity of their values (quantitative continuous or discrete, semi-quantitative or qualitative), their temporal variability (frequency and duration of sampling) and their topological structure (object with a geometry or not). We also observed changes in protocols and formats over the study period (2000-2010). All these data are localised and associated with geometric objects in the form of points, lines or polygons, making the structuring and interconnection of data complex.

All the data we collected were organised in one database, which was created specifically for our project, and was named the FresQueau Database. The conceptual model of the database was mainly based on the models of the source databases and on the recommendations of the French data administration service and water repositories, created in 1993 (SANDRE: http://www.sandre.eaufrance.fr). The five categories of data were centered on the sampling sites. The model was implemented in a PostgreSQL / PostGIS database, consisting in 81 tables including 13 tables from SANDRE. These tables were divided into the main categories mentioned above. The FresQueau Database had 60 million records and weighed twenty giga-octets.

2.5.2. Methodological approach

The challenge was to create an effective bridge between the two communities that, at first, do not expected the same results. The data scientists were attracted by

the huge quantity and the diversity of data, and interested in the design of new mining methods to answer well specified questions. Hydro-scientists hoped for methods to find not only all expected relationships but also novel and potentially useful ones. At the beginning of FresQueau project, we decided to develop an effective working approach centered on circumscribed questions defined by the hydro-scientists, in order to (1) manage the high heterogeneity and the complexity of the collected water data, (2) design the most suitable data mining method for each question, (3) respect the limited time of the project (three and a half years).

Evaluating water quality and the status of rivers is complex, and hydro-scientists had many expectations. After discussions with all the members of the FresQueau consortium –data scientists and water hydro-scientists-- eight questions were finally selected (appendix 1). We present three of them in the training cases (paragraph 3).

This approach did not include data collection and data modelling in the FresQueau database. Our approach included five steps: 1) design, 2) choice of the method, 3) data preparation, 4) application, 5) results and validation by the hydro-scientists. Figure 7 shows this approach.

Step 1. Design: in this step, hydro-scientists proposed one circumscribed question, explained their interest and the expected results. The consortium discussed these expectations. Data scientists formalised it with respect to general methods. They took into account spatial and temporal dependencies of the data. Finally, if all the hydro-scientists and data scientists agreed on the interest of the question, it was retained. The hydro-scientists then identified the categories of data concerned by the question.

Step 2. Choice of the method: the data scientists chose the data mining method best suited to the question. The difficulty of the question and the specificity of the data raised some new issues and encouraged the data scientists to innovate: they could design and implement a dedicated algorithm for the question and its associated data.

Step 3. Preparation of the data: the hydro-scientists selected the data and described their characteristics. In collaboration with the data scientists, they evaluated data availability and data quality (mainly completeness). Depending on the method selected, the hydro-scientists and the data scientists prepared the data, for instance, gathering some, discretizing others (i.e. transforming numerical variables into ordinal variables, using quantiles or thresholds of quality classes).

Step 4. Filtering and ranking the results: the selected algorithm was applied to the prepared data and produced results. The data scientists had to (1) explain to hydro-scientists how to read the results so that hydro-scientists could understand and interpret them; (2) implement specific, and often innovative discriminating filters to help the hydro-scientists rank the results.

Step 5. Validation and interpretation by the hydro-scientists: the hydro-scientists analysed the results. To validate the method and its application, the hydro-scientists used expected results. Finally, for the method to be convincing, the hydro-scientists also analysed unknown and/or surprising results which would lead to innovative interpretation.

Our approach was iterative. At the end of the first cycle, we refined (1) the question, and sometimes the expected answers, (2) the algorithm, (3) preparation of the data to improve data treatment and (4) the filters to produce (5) more accessible and clearer results, based on the hydro-scientists' expectations. If necessary, we started the cycle again.

We acquired a common language which facilitated the dialogue and made the collaboration progressively more efficient cycle after cycle.

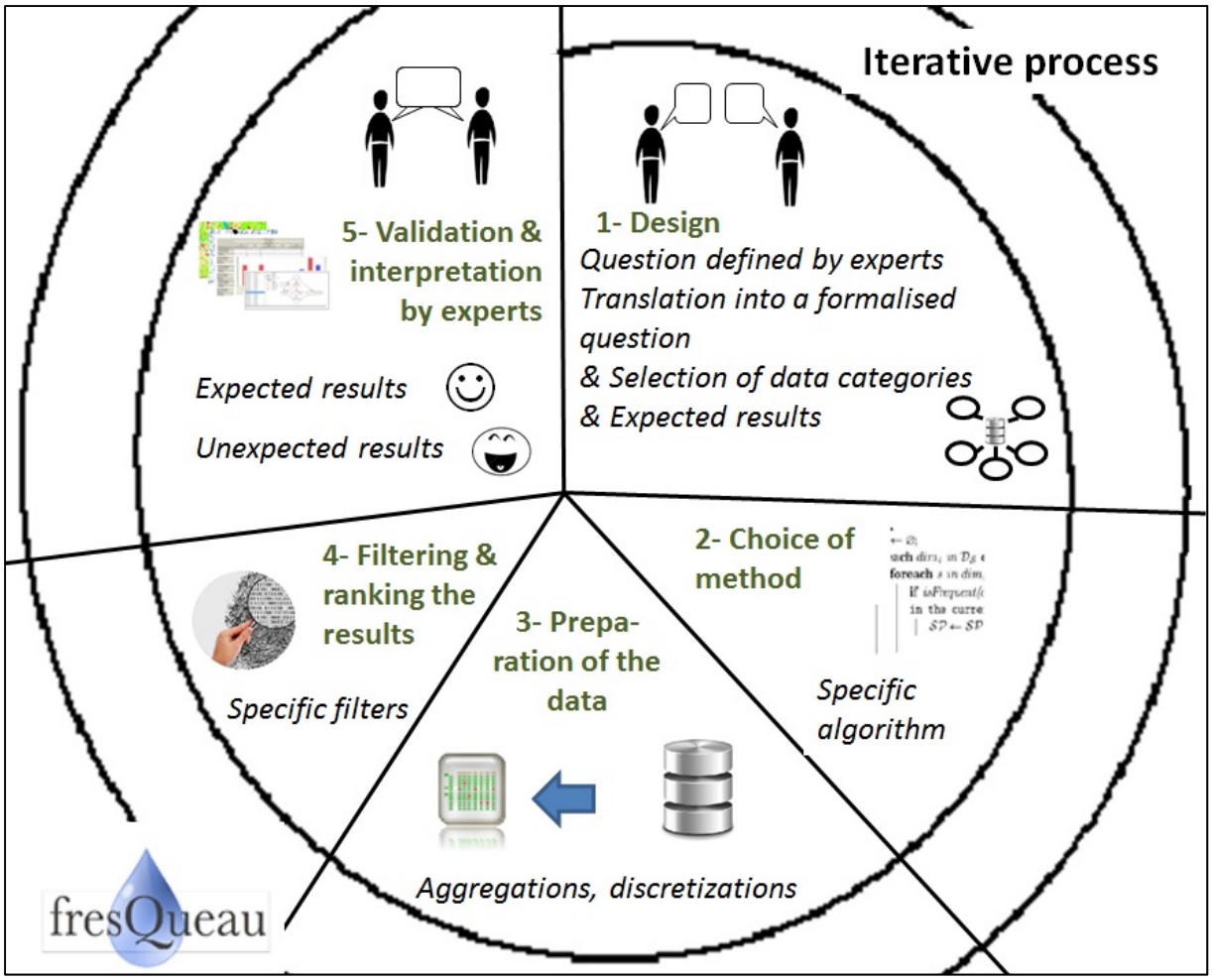


Figure 7: Diagram of FresQueau working approach

2.6 Three training cases

Here we present our working approach applied to three training cases. Each case can be seen as a recipe for using the considered method.

2.6.1. Training case 1: Discriminant temporal patterns to link physical-chemical parameters and biological groups in the assessment of hydro-ecosystems

Several human activities generate a combination of pressures which alter the abiotic components of the ecosystem, affect the biological communities and hence their ecological status (driving forces, pressure, state, impact, response, the DPSIR concept) (Kristensen, 2004)). Multi-stress conditions are multiscale (Dahm et al., 2013; Lalande, 2013). The issue facing the WFD is to have access to biological tools not only able to assess the ecological status but also to disentangle and identify the different pressures, to be able to propose the appropriate restoration action to achieve good ecological status (Feld et al., 2016b; Reyjol et al., 2014). Some authors (Larras et al., 2017; Mondy and Usseglio-Polatera, 2013; Villeneuve et al., 2015) have already tested the potential of supervised data mining methods on water quality data to reach these goals using different biological organisms. The search for predictive correlations using supervised learning requires the identification of explanatory and target variables among the descriptors originating from the description process described above. Once this step is completed, a large set of supervised learning techniques is available, including artificial neural networks, support vector machines, and decision trees. However, these off-the-shelf tools still need to be adapted to the nature of the available data as well as to the specific problem at hand. In our study domain, the most common methods include artificial neural networks (D'heygere et al., 2006; Dakou et al., 2006; Everaert et al., 2016; Tsai et al., 2016), and regression or classification trees (Džeroski, 2001; Feld et al., 2016a; Larras et al., 2017; Mondy and Usseglio-Polatera, 2013; Villeneuve et al., 2015). Furthermore, spatial and temporal dependencies require dedicated methods including Markov random fields, or bayesian networks (Adriaenssens et al., 2004;

Forio et al., 2016; Fytilis and Rizzo, 2013; Landuyt et al., 2016b; Van Looy et al., 2015).

State-of-the-art methods underline the importance of considering and combining biological and physical-chemical variables in order to discover relevant knowledge (Marzin et al., 2012). Yet none of the studies cited above accounted for temporal aspects as do temporal pattern mining approaches, which is the best way to analyze pollution and biological compartment dynamics.

2.6.1.1. Design

Among all possible pressures, we limited our study to the physical-chemical pressures of water. We did not address the problem of physical-chemical parameters of the sediments, nor hydromorphological parameters because of the scarcity of measurements in sediment or riparian indicators in the immediate vicinity of the sampling sites.

The question was: Can we link sets of physical-chemical parameters of water with bio-indices over time? The data categories concerned were the physical-chemical and biological river quality results and the location of the sampling sites.

The question was transformed by data scientists into a problem of temporal pattern mining in a sequence database, where each sequence links the successive physical-chemical and biological measurements made at one site. In addition, the mined patterns should be frequent and specific to a biological quality class. Five classes corresponded to five quality levels, illustrated by five colours: blue: high, green: good, yellow: moderate, orange: poor, and red: bad. For biological indices, we used the class thresholds given in the corresponding French norms (i.e. IBGN, IBD, IBMR, IPR op.cit. in 2). We decided to aggregate physical-chemical parameters in 12 pressure categories, as proposed in SEQ-eau (MEDD and AE, 2003) and listed in Table 1.

Table 1: List of physical-chemical pressure categories and their acronyms

Acronym	Pressure categories
MOOX	Oxidizable organic matter (e.g. O_2, DBO)
AZOT	Nitrogen matter, nitrate excluded
NITR	Nitrate
PHOS	Phosphorus matter
PAES	Suspended matter
TEMP	Temperature
ACID	Acidification parameters
MPMI	Heavy metals
PEST	Pesticides
HAP	Polycyclic aromatic hydrocarbons
PCB	Polychloro biphenyls
MPO	Other organic hydrocarbons

2.6.1.2. Choice of the method

The selected method consisted in extracting closed partially ordered patterns, (CPO-patterns) for each quality class. We therefore implemented the OrderSpan algorithm described by Fabrègue et al. (2015). A pattern was a sequence of items (or itemsets, e.g. ($AZOT_{Orange}$,$PHOS_{Yellow}$)) repeated in a sequence dataset (Srikant et Agrawal, 1996). In the case of CPO-patterns, a pattern appeared in a sequence if the order between all elements in the pattern was also observed in the sequence. Pattern mining methods usually provide a large set of results. Here we used various filtering processes to reduce the number of patterns and to provide relevant ones (Fabrègue et al., 2014).

2.6.1.3. Preparation of the data

As the method selected was qualitative, we needed to discretize the data. Physical-chemical and biological data extracted from the FresQueau database were modelled as follows: for each biological measure at a given site at time *t*, we considered the physical-chemical measurements made at the same site up to 6 or 4 months before *t*. The physical-chemical data were discretized according to SEQ-eau standard (MEDD and AE, 2003) and considered as items. The resulting values were modelled as sequences of itemsets (see Table 2) which wer linked to a biological quality class.

Table 2: Examples of sequences of physical-chemical values linked to two IBGN quality classes

Datasets	Sequences
IBGNBlue	⟨(AZOT^{Blue})(AZOT^{Green},PHOSBlue)⟩ ⟨(AZOT^{Blue},PHOS^{Green})(PHOS^{Green})(AZOT^{Yellow},PHOS^{Blue})⟩
IBGNOrange	⟨(AZOT^{Orange},PHOS^{Yellow})(AZOT^{Red},PHOS^{Orange})(AZOT^{Green},PHOS^{Yellow})⟩ ⟨(PHOS^{Orange})(AZOT^{Orange},PHOS^{Yellow})(AZOT^{Green})⟩

2.6.1.1. Filtering and ranking the results

The method was applied to all sites in the database for a period of 10 years, resulting in 8,900 sequences for three biological indices (IBGN, IPR, IBD). The results are lists of CPO-patterns for each quality class of biological parameters. Each CPO-pattern summarises a set of sequences. For example, the CPO-pattern in

Figure 8 (from Fabrègue et al., 2014) is supported by all sequences in the dataset IBGN^{Orange} and none in the dataset IBGN^{Blue}. Indeed, in all sequences in dataset IBGN^{Orange}, itemsets (AZOT^{Orange} PHOS^{Yellow}) and (PHOS^{Orange}) both appear and are followed by item AZOT^{Green}. Conversely, (AZOT^{Orange} PHOS^{Yellow}) and (PHOS^{Orange}) are ordered differently in the sequences.

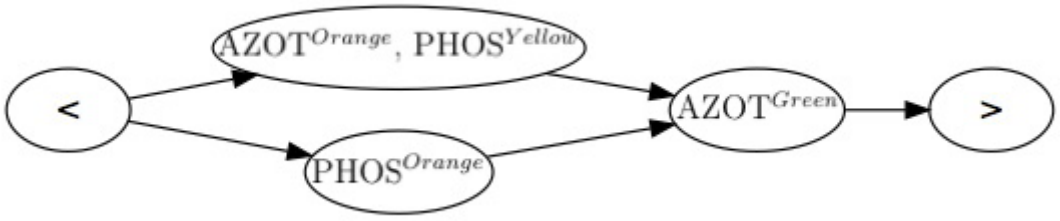


Figure 8: a CPO pattern extracted from the sequences in Table 2, for the target IBGN^{Orange}

The results were further characterised by three measures used as filters: frequency, discrimination and redundancy. Frequency allowed us to select the CPO-patterns. A first high frequency threshold was selected to remove the too obvious CPO-patterns. A second lower threshold was chosen to select the most frequent and

significant CPO-patterns in the dataset. These thresholds were chosen by the user. Discrimination made it possible to select CPO-patterns which were specific to a quality class dataset. Redundancy as used to select only one representative CPO-pattern among similar results. Fabrègue et al. (2014) detailed the computation of these filters and had combined them in a specific index: the Pattern Balance Index to select the most frequent, discriminant and non-redundant CPO-patterns.

2.6.1.5. Validation and interpretation by the hydro-scientists

The approach presented here is suitable for temporal datasets with multiple variables, in this case, biology and physical-chemistry. In addition, the knowledge extracted using temporal patterns is easy for hydro-scientists to analyse for a descriptive point of view.

As shown in Figure 8, among the expected results, CPO-patterns reveal good connections between physical-chemical quality and biological quality. A quite unexpected result was to find, most of the time, a systematic shift of one quality class between biology, always worse, and physical-chemistry. For example, in Figure 9, all the physical-chemical parameters indicating good quality are in green and the target, IBGN, of moderate quality is in yellow. This result confirms that the integrated information contained in biological indices is more sensitive to pollutants than the information built on physico-chemical parameters.

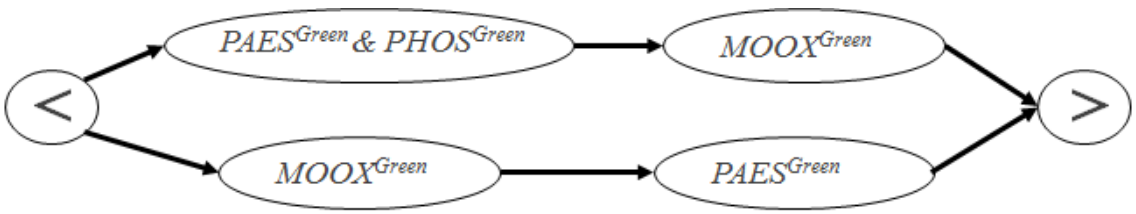


Figure 9: CPO pattern supported by all sequences in dataset IBGN^{yellow}

For some indices, the hydro-scientists expected specific pressure categories. As expected, the most frequent pressure category in CPO-patterns concerning IBD was phosphorous. On the other hand, we did not find the pressure category nitrogen, which could have been expected for IBD. We assumed that this pressure category was not sufficiently discriminant in our dataset.

2.6.2. *Training case 2: Relational concept analysis to search for links between macroinvertebrate life traits and the status of the hydro-ecosystem*

Life traits describe how living organisms fulfill their biological functions (e.g. feeding and reproduction) and what their ecological preferences are (e.g. saprobic or oligotrophic conditions). According to Statzner and Bêche (2010) prior trait responses can be used to assess the status of an ecosystem: for example, the form of respiration depends on organic pollution.

Analyses of (1) the life traits of the biological part of the ecosystem, (2) the links between the hydromorphological, physical-chemical and biological parts, and between these parts and pressures caused by humans or by environmental conditions, are promising alternative ways to conduct a comprehensive diagnosis (Archaimbault et al., 2010; Dolédec and Statzner, 2008; Usseglio-Polatera et al., 2000) even in a multi-stress environment (Mondy and Usseglio-Polatera, 2013), or to make predictions (Feio and Dolédec, 2012). De Lange et al. (2010) emphasized the advantages of this type of predictive diagnosis for toxic pollution. Blanck (2002) suggested monitoring community evolution in a context of chronic contamination using the pollution induced community tolerance (PICT) approach.

2.6.2.1. Design

Our original question was: Are there any links between variations in an environmental (physical, physical-chemical) parameter and variations in a biological trait? The question was narrowed down to “Are there any links between variations in

a physical, physical-chemical environmental parameter and the frequency of expression of a given modality of a given biological trait?" In the following, we use the term "trait" to refer to "the frequency of expression of a given modality of a given biological trait". The data categories concerned were the physical-chemical and biological results, river quality results, and the location of the sampling sites. The question was transformed by data scientists into a problem of extracting implication rules from relational data. Such rules were formalised as: "if premise then conclusion".

2.6.2.2. Choice of the method

The method chosen was relational concept analysis (RCA), a method for exploring relational data (Dolques et al., 2016b) and discovering rules (Dolques et al., 2016a). RCA iteratively computes several concept lattices connected by links (relational attributes) which represent the relations between objects. Rules are extracted by considering a main lattice (which contains rule premises) and the relational attributes that connect this lattice to the others.

2.6.2.3. Preparation of the data

The method selected was qualitative, so we had to discretize the data. For each site and year, the average annual measurements of each physical-chemical parameter, each physical parameter and each macroinvertebrate set were selected. Taxa were associated with their traits. Numerical values were separated into several quantiles: 1) three quantiles for the relative abundances of taxa: S>25%, S>50%, S>75%, respectively for low, high and very high abundances; 3 quantiles for affinities: Q1, Q2, Q3 for low, medium and strong affinities; five quantiles for physical and physical-chemical parameters for high, good, moderate, poor and bad qualities: R1, R2, R3, R4, R5. We defined a binary table for each object dataset (physical-chemical parameters, sites, macroinvertebrates, traits) and a binary table to represent each relational level (e.g. affinity level Q2 for traits of macroinvertebrates). The corresponding relational schema is presented in Figure 10.

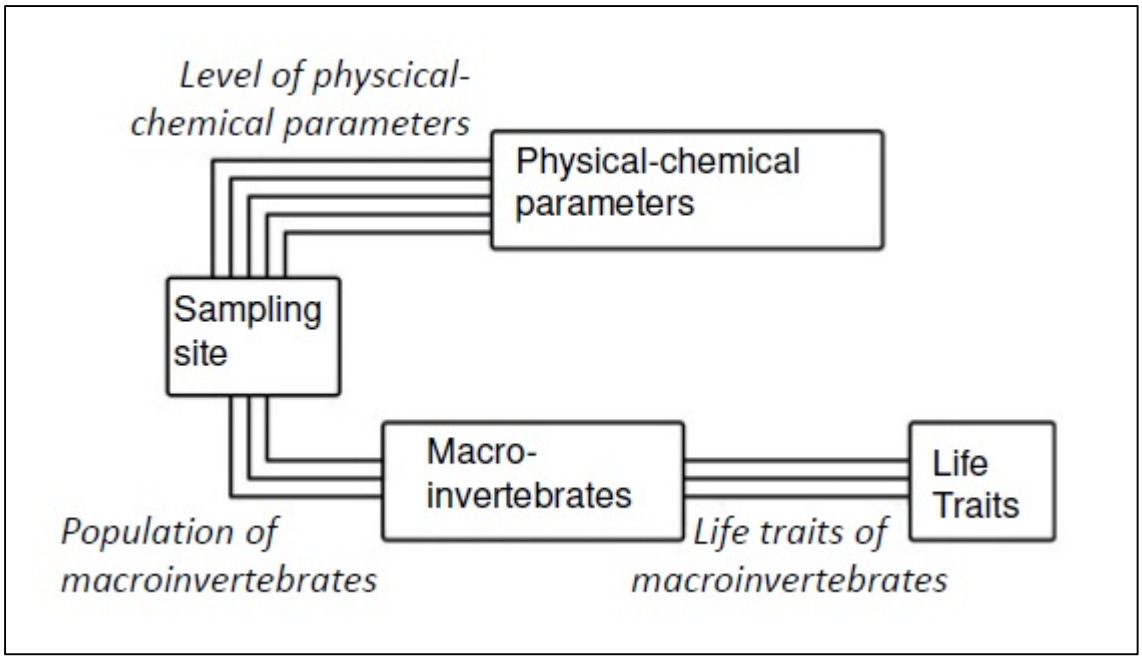


Figure 10: The relational schema between physical-chemical parameters and traits of macroinvertebrates living at a given site (rectangles represent tables, lines represent the number of quantiles used to rank the results in the different tables: 5 quantiles for physical-chemical parameters, 3 quantiles for macroinvertebrate abundances and 3 quantiles for trait affinities).

2.6.2.4. Filtering and ranking the results

We limited the experiment to a regional subset (Alsace in Rhine-Meuse) with 49 sampling sites. The first table, called R, contains 27 physical and physical-chemical parameters collected at the 49 stream sites. The second table, L, gives the population abundance for 197 macroinvertebrates collected at the same 49 sites. The third table, Q, describes 18 different traits, each having several modalities (e.g. for the trait "life cycle", there are two possible modalities: "less than a year" or "more than a year"). Three levels of affinity –low, medium and high – link each taxon of table L to each modality of each trait in table Q. In total, we had 113 combinations of trait-modality and 339 combinations of trait-modality-affinity.

The results are rules which are ranked by different weighting of their premises or conclusions. The first rules, with highest weights, which are evident and valid everywhere are eliminated, such as "all taxa prefer fresh water". An example of an extracted rule (from Dolques et al. 2016a) is given below (Rule 1): "if at a given site, more than half of its highly represented taxa are transversally distributed mainly on

banks and side-arms, then the state of its banks is considered as good". This rule was supported by 15% of sites in the dataset. This percentage may seem low but is high compared to the many possible combinations trait-modality-affinity: it is the highest percentage found. Indeed the presence of taxa living on banks or side-arms is only possible at sites where the banks are conserved or restored.

Rule 1:

S>50% high_taxon_abundance (∃ Q3 strong_affinity (transversal distribution : bank sand side-arms))

→ ∃ R2 good_state (banks)

Two sorts of rules can be extracted: rules with physical-chemical parameters in the premise or conversely, rules with traits in the premise.

2.6.2.5. Validation and interpretation by the hydro-scientists

We retained the 189 first rules on the 1428 generated. These 189 rules were all supported by a small percentage of sites: 15% was the highest one, which is relatively important for the large number of possible combinations "trait-modality-affinity". But on account of our reduced dataset (49 sampling sites), our results are linked to a specific context.

Among the highest weighted results, we found some expected results, as in Rule 2, that means: "if numerous taxa have a strong affinity for longitudinal distribution metapotamon, then the nitrite concentration results in a moderate quality class". According to the system of longitudinal zonation defined by Illies and Botosaneanu (1963), metapotamon is the last reach of a river before the estuary where saprobic and trophic levels are the highest and consequently, so is the concentration of nitrites. Actually, the gradient of nitrites in our dataset was important (minimum: 0,01 mg/L, maximum: 1,54 mg/L; average: 0,14 mg/L, standard deviation: 0,202: which justifies that the rule has a low support.

Rule 2:

S>50% high_taxon_abundance (∃ Q2 medium_affinity (longitudinal distribution : metapotamon))

→ ∃ R3 medium state (NO_2^-)

Besides, many rules, among the ones with the highest supports, concerned physical parameters, as shown by Rule 3: "if numerous taxa have a strong affinity for slow flow, then the state of hydrology is poor" and Rule 4: "if few taxa have a medium affinity for fast flow, then the state of hydrology of the minor bed and the connectivity with side-arms are poor". These rules were supported by 8% of the sites in the dataset. This result could seem unexpected but could be explained by the characteristics of our study area, the upper Rhine. Indeed, the main channel of the river was channelized in the middle of the 19th century to protect against flooding and to promote human activities such as navigation, stabilising the banks to increase agricultural land, and the production of hydroelectricity (Arnaud et al., 2015). As a consequence, the Rhine lost lateral connectivity, its side-channels which were disconnected are now only fed by groundwater (Meyer et al., 2013) and has a low flow. Some of our sampling sites were located on these disconnected side-channels: their physical status was degraded, especially their hydrological status, due to the lack of significant hydraulic feed, and connectivity due to the now oversized bed.

Rule 3:

S>50% high_taxon_abundance (∃ Q3 strong_affinity (flow : slow))

→ ∃ R4 poor state (hydrology)

Rule 4 :

S>50% low_taxon_abundance (∃ Q2 medium_affinity (flow : fast))

→ ∃ R4 poor state (hydrology, minor bed, connectivity)

2.6.3. Training case 3: Multiscale spatial modelling to link land uses with the ecological status of the hydro-ecosystem

The WFD requires "the identification of the significant anthropological pressures and the evaluation of their impacts on the ecological status of streams " (Appendix II) (European Council, 2000). In fact, different human activities (urbanization, farming, industry, transport, navigation, sports and recreation, etc.) create pressures on surface waters and hydro-ecosystems. There are many such pressures and they are all the more difficult to qualify and quantify in large study areas. It is thus necessary to identify and rank the causes of physical, biogeochemical and ecological changes in the functional processes of the aquatic ecosystems for diagnosis, decisions concerning restoration or conservation actions, or for assessment of the efficiency of actions undertaken.

Surface waters are very vulnerable to diffuse or point pollution as well as to water withdrawals and morphological changes. Although some industrial, domestic or urban pressures can be identified directly from data on existing effluents in the environment, this is not the case for diffuse pollution irrespective of its origin. Diffuse pressures come from surface sources and their mode of transfer is inevitably spatialized. Data on land use and the resulting indicators are usually accepted as reliable indicators of the physical and biological integrity of aquatic ecosystems. Furthermore, data on land use are defined in space, and are thus reasonably homogeneous and easily available. They are a good tool to estimate diffuse pressures on a basis of a given land use. We shall refer to numerous studies connecting land use stressors and the biological, chemical and physical quality of running waters which can be defined by observations or indices (Allan, 2004a, 2004b; Bruno et al., 2014; Johnson and Host, 2010; Kail et al., 2015; Meador and Goldstein, 2003; Waite et al., 2014). Nevertheless, the models we present require further development both to improve knowledge of these complex processes and to design operational tools.

2.6.3.1. Design

Our initial question was: What are the relations between pressures, chemical parameters and biological indices? To identify the impacts of multi-stressors, we chose to limit the question to: what are the relations between land use stressors and the biological index based on macroinvertebrates: IBGN (see paragraph 2.1)? The distribution of macroinvertebrate communities is influenced by two main types of processes:

- ”spatially structured ecological processes and neighbourhood interactions such as larval dispersal, migration, competition and the spread of disease” (Hamylton, 2013);
- external physical-chemical processes structured and generated by the surrounding anthropological activities.

Running waters depend on structural controls at catchment scale, reach scale, channel pattern differences and microscale variations in channel bed forms, all of which vary over different time scales (Friberg et al., 2016). As (Allan, 2004b), we considered three scales: macro- (catchment), meso- (reach) and micro-scale (sampling site) (Lalande et al., 2014). All these processes give rise to spatial dependences in macroinvertebrate communities and in the pressures to which they are exposed.

The question was transformed by data scientists into a problem of spatial modelling. The appearance of geographical information systems (GIS) notably contributed to analyses and studies of spatialized data. In the FresQueau database, thanks to PostGis (see 2.1), these spatialized data were available. However, it should be noted that GIS was developed independently of the statistical analysis of the spatial data. The integration of the statistical tools in the GIS software is beginning to improve, but the proposed statistical tools themselves are not really suited to the regionalised spatialized nature of the variables because they do not account for the spatial autocorrelation of the values (Anselin, 1995; Aubry and Piegay, 2001; Lichstein et al., 2002).

By working on the spatial interactions between land uses and the quality of streams, three additional difficulties emerged:

- the spatial structure linked to by the river network (included upstream/downstream relations);
- the number of water quality stations;
- the sparse and irregular dataset after preprocessing we had at our disposal.

The data categories concerned were macroinvertebrate indices (river quality), location of the sampling sites (sampling sites), hydrographic network data and land use data.

2.6.3.2. Choice of the method

The method chosen was based on statistics. Spatial correlation methods were used to detect spatial dependency which happens between two water quality stations, due to river network characteristics, by analyzing distance or neighborhood effects. Two different kinds of spatial correlation tools were used: graphic tools, such as the decay distance, or statistical tests, such as Moran (1950), Geary (1954) indices for neighbouring effects and Mantel (1967) for distance effects. We also used a specific test developed by Lalande (2013) able to integrate orientation and network constraints. According to the results of this first series of tests, two different approaches could be used to model relationships between water quality indices and land use parameters. When spatial dependency between two stations was not significant, water quality data was considered to be independent among sampling sites. In this case, considering the small number of observations, Spearman's correlation coefficient could be calculated between the median of a given water quality index and land use and riparian parameters. When network effect analysis revealed a spatial dependence among stations, auto-regressive modelling could be used to model land use and water quality. In these cases, water quality at a given site depended on the quality of water and land use conditions at the next site upstream (Lalande, 2013).

2.6.3.3. Preparation of the data

The method chosen was quantitative. We limited our experiment to two tributaries of the River Saone: the Azergues and the Ognon.

A total of 85 pressure stressors were built for each of the sampling sites in the Corine Land Cover 2006 (SOeS-EEA, http://www.eea.europa.eu/publications/COR0-landcover) database at macro-scale, and from a dedicated land use map for riparian areas (Decherf and al. 2011) at meso- and micro-scales.

The IBGN was used to qualify the ecological status of the streams. Spatial analysis was performed on the year richest in data: 2004 for Azergues and 2008 for Ognon, with respectively 22 and 10 sites sampled in IBGN, (Figure 11). The average distance between two successive water quality sites was 5 km for Azergues and 28.8 km for Ognon. The topology of the hydrographic networks (BD Topo ® IGN, http://www.professionnels.ign.fr/bdtopo) was checked in order to verify connection and upstream/downstream orientation. We highlighted 130 spatial interactions for Azergues and 31 for the Ognon. From these interactions, we built neighbouring and distance matrices.

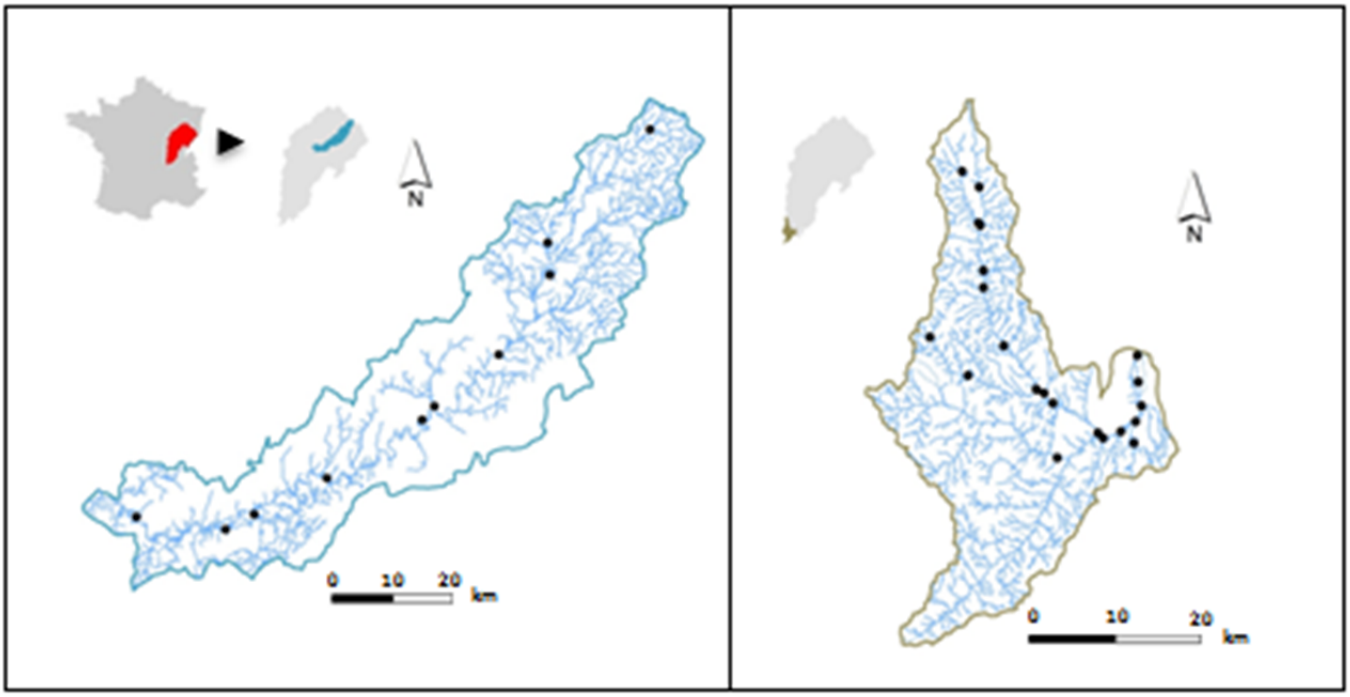


Figure 11: Hydrographic network and IBGN sampling sites in the Ognon (left) and Azergues (right) watersheds, two tributaries of the River Saone (France) sampled in 2008 and 2004, respectively

2.6.3.4. Filtering and ranking the results

Distance-decay analysis showed that the quality of water generally decreased from the upstream to downstream, which was predictable. Four tests, described in part 3.3.2, were applied to the series of status indicators for both experimental zones to check for significant autocorrelations. Spatial correlations were detected for the Azergues watershed whereas no spatial correlation was found for the Ognon watershed.

The autoregressive model for Azergues basin provided, for a p-value inferior to 5%, the identification of 17 relevant indicators, among which the presence of forest on the whole reach upstream of the sampling site, meadow zones in a 300 m wide band on both sides of the upstream river system, and the presence of an artificial zone near the water quality sampling site. Spearman's coefficients calculated for the Ognon basin provided, for a p-value inferior to 5%, the identification of 15 pressure stressors including the presence of agricultural zones and forest near the water quality sampling site. These results are summarized in Table 3. Finally, twenty-eight pressure stressors showed significant correlations between land use and IBGN for a very p-value.

Table 3: Land use indicators proposed for the three scales – macro for all the watershed, meso for around the reach and micro for around the sampling site- and correlations found with IBGN

(Legend: 1) empty and white cell: no land use indicator (LUI); 2) white cell with X: a LUI calculated but no correlation found with IBGN; 3) grey cell with letter(s): a LUI calculated and a statistically significant correlation (for a p-value inferior to 5%) found with IBGN, in the Ognon watershed (O), in the Azergues watershed (A), in the two watersheds (A&O))

	Macro-scale	Meso-scale					Micro-scale				
		Width of the buffer zones around the reach (m)					Width of the buffer zones around the sampling site (m)				
Land Use		**10**	**30**	**50**	**100**	**300**	**10**	**30**	**50**	**100**	**300**
Natural vegetation		O	x	x	x		x	x	x	O	
Forest	A&O	A&O	x	x	x		O	O	x	O	
Herbaceous vegetation		X	x	x	x		x	x	x	x	
Agricultural area	A&O		O	x	x	x		O	O	O	O
Arable land	O		A	A	A&O	A		A	A	A	A
Vineyard			x	A	x	x		x	x	x	x
Grassland			x	x	A	A		x	x	x	x
Artificial area	x		x	x	x			A	A	A	
Impervious area	x		x	x	x			x	x	x	
Non impervious & artificial area			x	x	x			x	x	x	
Road			x	x	x			x	x	x	

2.6.3.5. Validation and interpretation by the hydro-scientists

As expected, the correlations were positive for natural and semi-natural land uses (forest) as well as for grasslands, and negative for artificial types of land use (urban area and roads) and agricultural land uses. It should be noted that it is not possible to build general indicators of the pressures exerted by the land use on the water quality: it seems very dependent on the study cases. Nevertheless, the results confirmed previous studies (Allan, 2004a; Meador and Goldstein, 2003). Therefore, in this kind of area, main land uses (forest and agricultural area) give indication on upstream downstream anthropogenisation gradients for the basin scale. Moreover, at

the meso-scale, when we took the upstream network into account, the presence of forest in a 10m-width corridor and of arable land in a 100m-width corridor played a preponderant role. Always as expected, when we analyzed the results with the descriptive statistics of land use stressors, we noticed that the significant indicators were those which were the most variable or/and the most important. Even if no indicator at the micro-scale was significant for both catchments, we noticed direct relations between macroinvertebrate habitat and land use as forest or agricultural area for Ognon catchment, and arable land and artificial land for Azergues catchment. So, the three scales are necessary for the analysis of pressures. Litterature (Johnson and Host, 2010; Osborne and Kovacic, 1993) indicates a lot of widths for meso- and micro- scales, one of our contribution is that the width of the right-of-way corridor on which the indicators are constructed had little influence in the values of the pressure indicators whatever the type of land use concerned. Thus, the number of meso indicators could be limited by considering only one or two rights-of-way per type of land use. Finally, the spatial dependences in a succession of upstream / downstream sites are not systematic and are connected to the density of sample points.

The spatial structure of the relations between the data was determined by the river network. The river network is a connected and directed entity, which was treated on a hierarchical basis. Before any calculation or analysis, a geographic information system was used to check and correct the topology of the geographic layer containing hydrographic networks. We decided to favour the spatial integrity of the hydrographic network rather than the volume of data. Indeed, spatial constraints were still difficult to take into account from a statistical point of view. Moreover, in our case, a supplementary constraint was the scarcity of data available for our test. But despite these constraints and in addition to identifying pressure indicators, the study of two small areas allowed us to detect two kind of anomalies due to: (i) data characteristics such as the accuracy of the water quality indices and of the land use map, (ii) local characteristics such as the proximity of a confluence of a tributary, or a point pollution source.

Scaling up the test conducted in the Azergues and Ognon watersheds to all the sampling sites on the River Saone would confirm or invalidate the main conclusions we drew: the role of the width of the upstream network indicators, the

presence of artificial land use in micro indicators, the importance of small tributaries, etc. The resulting increase in the number of data would also require the development of other methods.

2.7 Discussion

2.7.1. Availability and quality of data

The application of WFD increased the number of assessments of European running waters. In France, water data became "big data". These data are public. National formats and references have existed since 1993 in the SANDRE (op. cit.) and continue to evolve. However, there is always considerable variability in the sources, often depending on dataset owners' objectives which may differ in time and space. Moreover SANDRE does not include all the data on water quality: for example, land use data. Collecting and structuring these data covering one third of metropolitan France for the FresQueau project was difficult and time consuming, it accounted for more than one third of the project in human resources, and multiple exchanges between hydro-scientists and data scientists. As it was aware of these difficulties, in 2016, the French Biodiversity Agency (French acronym AFB) introduced an open data service named Hub'Eau (http://www.hubeau.fr). Hub'Eau has begun to provide open data on groundwater, waste water treatment, fish species in running waters and recently, in our domain: the quality of running waters.

Studies on heterogeneous data are scarce because such studies are not easy (Tsai et al., 2016). In the FresQueau project, Berrahou et al. (2015) specifically analysed the quality of the data, including their thematic and temporal accuracy, and the logical coherence, consistency and completeness of the data collected. One of the major quality problems was the completeness of certain types of data. The absence of some data could prevent the identification of relationships between objects in the FresQueau project database. We faced two major completeness problems. The first one was that out of the 11,329 data collected at the sampling sites, 28% were not connected to a river. We reduced this rate to 10% by applying

several geographical queries on sampling sites without a recorded river. The second problem was the lack of results for some minor physical-chemical elements. For instance, in the Rhine-Meuse watershed, filling rates of physical-chemical results (river quality data) were 89% for major elements (33 parameters, e.g. nitrate) but only 30% for minor elements (611 parameters, e.g. heavy metals or pesticides). The lack of results for minor elements explains why we failed to find them either in the CPO-patterns in the first case or in the rules in the second case.

2.7.2. Innovation and performance of the approach

Aiming to describe the ecological status of running water using available public data, we developed three specific data mining methods which gave expected results, allowing to validate these methods, and also revealed promising unexpected results. These methods were rapid: for instance, in the first case, after pre-processing and transformation, less than two minutes were needed to generate CPO-patterns on the whole FresQueau dataset (Fabrègue et al., 2014). The specific filters allowed us to select the most frequent, discriminant and non-redundant results - e.g. the PB-index (Pattern Balance Index) (Fabrègue et al., 2014) in the first case -, ordered and limited the number of these results –e.g. in the second case, 189 rules were retained out of a total of 1,428 rules generated (Dolques et al., 2016a). We demonstrated the efficiency of CPO-patterns (case 1) to explore the temporal abiotic trajectories leading to a given biological status, the potential of relational concept analysis (RCA) (case 2) to consider multi-relationships between abiotic parameters, taxa and traits of these taxa, or spatial correlation methods (case 3) to account for spatial dependences.

Some difficulties remain: we were unable to solve the shift to large scale for the second and third methods. We applied RCA on a reduced dataset (49 sampling sites). On larger datasets, the size of concept lattices explodes and it becomes difficult to select pertinent rules, if we want to be exhausive. Interestingness measures need to be designed and implemented to select pertinent rules and possibly to compute only a part of the lattices. In case 3, the difficulty was to create land use indicators for meso- and micro-scales and to account for the upstream/downstream relationships in the hydrographic network. This difficulty is the

reason why, in this case, we had to use statistical methods rather than data mining methods. Besides, a statistical test on the significance of the results discovered by data mining methods (e.g. by comparing the results obtained with those obtained from the randomized data) has to be done, but seems difficult to achieve because of the volume of data.

Our approach is based on a continuous dialogue between data scientists and hydro-scientists, thus improving the KDD process. By including the elaboration of a common design at the beginning of the process and ongoing discussions between hydro-scientists and data scientists throughout the process, our approach allowed us to choose and specifically adapt data mining methods, and to obtain interesting results on a complex system based on a large volume of complex (varied, heterogeneous, multi-source) data. We recommend iteratively beginning the entire process again when necessary. As reported by Tsai et al. (2016), our approach makes it possible to explore complex relationships between heterogeneous data in the field of natural water quality.

2.7.3. Interdisciplinary approach

According to Legay (1999), environmental issues are complex and at least require multi-disciplinary approaches. For instance, the different biological group responses could have weak concordance as showed by Larsen et al. (2012). Our approach is a real interdisciplinary approach, rather than a multi-disciplinary approach in which the disciplines establish their own conceptual and methodological choices side by side rather than as a result of collaboration (Rodela and Alasevic, 2017). In a context of global change, interdisciplinary research is critical for decision makers (Reid and Mooney, 2016). Thanks to their experience in modelling adaptation measures in a river basin (Orb, France) in a global change scenario, Girard et al. (2015), concluded that the interdisciplinary approach is "by no means a trivial task". Like them, during our research, we implemented a continuous dialogue to acquire a common language and shared concepts and progressively refined our approach in an iterative process.

According to Pohl (2005), researchers need several years of collaboration (at least six years in his case studies) to become acquainted with the other "culture" and to be able to produce together. Our approach allowed us to establish shared scientific questions during the design step and to solve them more efficiently. We worked together on the FresQueau project for only four years, but for much longer if we include our previous project (Grac et al., 2006) and the work of Bertaux (2010). More time is needed to exploit the promising results we obtained with the three chosen data mining methods. As reported by Statzner and Bêche (2010): "the field still has a long way to go to deliver on the promised objective: the reliable resolution of multiple stressor effects on running water ecosystems."

2.8 Conclusion

In many domains, "big data" is seen as a new "Eldorado" for data miners to discover new knowledge. And in practice, data mining methods, and in particular, readable methods, do enable the exploration of heterogeneous data with complex relationships and do facilitate dialogue between data scientists and hydro-scientists. Our experience shows that permanent interdisciplinary dialogue between data scientists and domain experts, as part of an iterative process, is needed to produce innovative validated results. In the case of running waters, the first difficulties are collecting and modelling very different data and ensuring their quality (e.g. completeness). In the case of biological and physical-chemical data, numerous indicators exist at different levels (e.g. biological indices, physical-chemical pressure categories, quality classes) which can be explored. But in the case of physical or land use data, before the data can be mined, relevant indicators need to be produced from the raw data, which is exactly what we propose in our third case.

2.9 Acknowledgments

This work was funded by the French National Research Agency (ANR), as part of the ANR11_MONU14 FresQueau project. We would like to thank D. Levet, retired from Aquascop firm, for her contribution to the project. The first author would

like to personally thank Franck Chauveau and Jean-Nicolas Beisel for their support and advice.

APPENDIX 1: the eight questions of hydro-scientists of FresQueau consortium

1. Are there any links between variations of a given physical, physical-chemical parameter and the occurence of a given modality of a given biological trait?"
2. Do physical-chemical parameters influence downstream bio-indices?
3. Do the physical-chemical parameters cause the absence or presence of taxa or changes in taxa abundance?
4. What are the links between bio-indices and hydromorphological characteristics?
5. Can we define a typology of sampling sites based on environmental context (driving data and hydrographic network)?
6. What are the relationships between pressures (land use stressors), chemical parameters and bio- indices?
7. Can we detect upstream/downstream relations between pressures (land use stressors), physical-chemical parameters and bio-indices?
8. What links between bio-indices and hydrology can be highlighted?

CHAPITRE III : Fouiller les séquences de qualité physico-chimique précédant un état biologique d'une masse d'eau

1 Résumé élargi

Ce chapitre a été soumis, sous le titre *« Mining the sequential patterns of water quality preceding biological status of waterbodies »,* à la revue *Ecological Indicators* le 16 juin 2019, qui l'a refusé le 22 août 2019 dans le format actuel jugé trop long, malgré « *son intérêt et son aspect innovant* », L'article est en cours de révision pour le soumettre dans une version plus courte à *Ecological Indicators.* Nous avons choisi de présenter ici la version longue initiale, jugeant que le développement de la méthode proposée a son intérêt et toute sa place dans le cadre de cette thèse.

La collaboration menée entre informaticiens et hydro-écologues nous a permis de tester l'adaptation de trois méthodes de fouilles de données aux questions relatives à l'évaluation de l'état des masses d'eau, sur des données recueillies sur l'Est de la France. Deux de ces trois méthodes sont non supervisées, dont la méthode d'extraction de motifs séquentiels partiellement ordonnés qui a été utilisée pour répondre à la question « les successions de mesures physico-chimiques sur plusieurs mois à années permettent-elles d'expliquer les réponses des différents indices biologiques à un instant donné ? ». Ici les motifs séquentiels sont partiellement ordonnés dans le temps. Nous les dénommerons motifs ci-après.

L'objectif de ce chapitre est d'investiguer si certaines des successions peuvent être identifiées comme caractéristiques de réponses types. Si tel était le cas, ces successions caractéristiques pourraient servir à prédire une réponse biologique. Pour cela, nous avons étendu le terrain d'étude à l'ensemble des données de surveillance de la France métropolitaine réalisées sur 1 781 stations, pour la période 2007-2013.

Nous avons créé une base de données spécifique pour accueillir les données de surveillance dont nous disposions. Nous avons optimisé l'extraction des motifs en développant un nouvel algorithme, PRESTOR, directement relié à la base de données. Les motifs sont des successions de mesures d'altérations physico-

chimiques discrétisées en classes de qualité ou d'état ayant précédées un indice biologique donné – IBGN, I2M2, IBD, IBMR ou IPR – dans un état donné.

Le vocabulaire associé au domaine de la fouille de données est très spécifique et nous utilisons par la suite celui dédié aux motifs (Figure 12). L'indice biologique donné dans un état donné est le contexte du motif. Les motifs sont recherchés dans les séquences, c'est-à-dire les successions de mesures d'altérations physico-chimiques discrétisées en classes ayant précédé le contexte et mesurées sur un ensemble de stations, chacune à des dates données. Dans un motif, chaque altération dans une classe donnée est un item, c'est-à-dire un événement. Un itemset correspond à un ensemble d'événements ayant eu lieu en même temps : il peut être donc composé de un ou plusieurs items. Une bifurcation dans un motif signifie que celui-ci est la synthèse graphique de deux séquences possibles : la série d'itemsets située en haut de la bifurcation peut avoir précédé, ou à l'inverse succédé à la série d'itemsets du bas (

Figure **12**). La fréquence d'un motif est le rapport entre le nombre de séquences vérifiant ce motif et l'ensemble des séquences ayant le contexte choisi (un mauvais IBGN dans l'exemple ci-dessous). Le nombre de séquences vérifiant le motif est appelé support du motif.

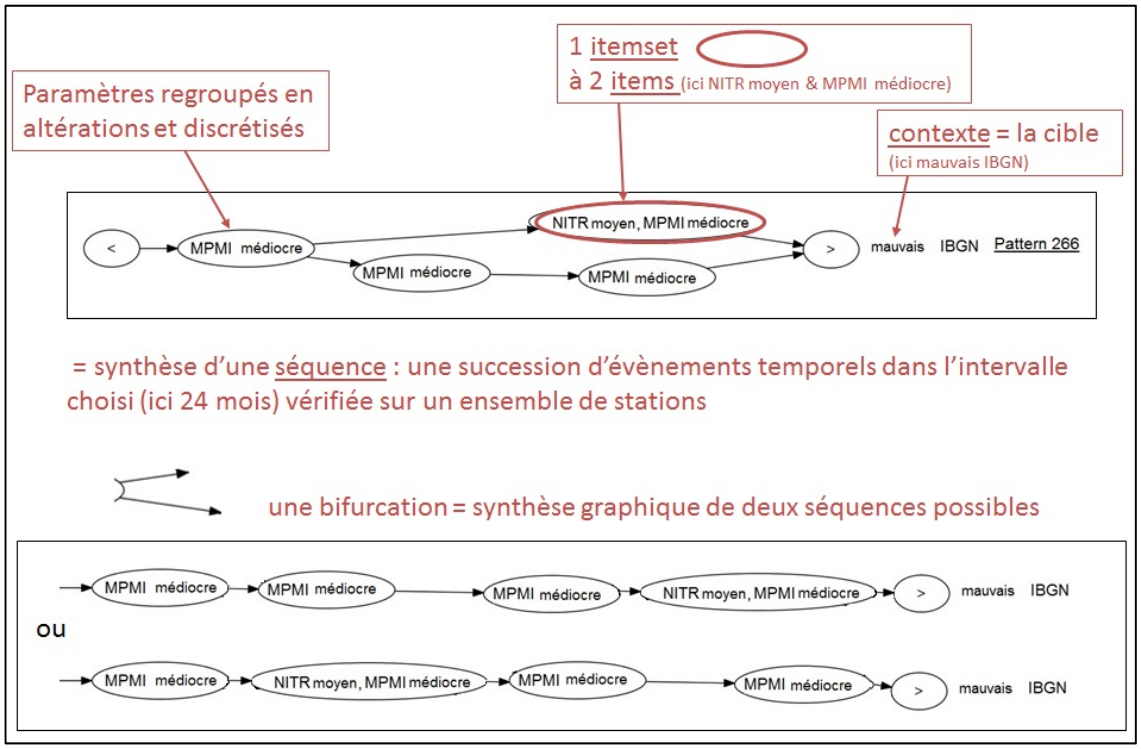


Figure 12 : illustration d'un motif et notions de vocabulaire associé (mots soulignés)

PRESTOR permet à l'opérateur de choisir plusieurs critères avant l'extraction :

- la zone géographique à considérer : la France entière ou une hydro-éco-région,
- une période comprise entre 2007 et 2013 ou la période complète,
- les grilles de valeurs-seuils à appliquer aux données physico-chimiques : celles du SEQ-eau (Système d'Evaluation de la Qualité de l'Eau) (MEDD and AE, 2003) ou les grilles DCE 2012 (MEEM, 2012),
- la durée des successions temporelles à extraire,
- la fréquence minimale en dessous de laquelle l'extraction de motifs doit s'arrêter ; la fréquence d'un motif est le ratio entre les séquences temporelles vérifiant le motif et l'ensemble des séquences temporelles disponibles,
- les éventuelles altérations à ne pas prendre en compte.

Les extractions de motifs avec PRESTOR sont rapides et stables : une extraction recommencée avec les mêmes critères donne les mêmes résultats. Le nombre de motifs obtenu augmente avec la taille de la zone géographique, la période utilisée, la durée des successions temporelles à extraire. A l'inverse, il diminue si la fréquence minimale est augmentée. Mais il n'y a pas de relation proportionnelle entre le nombre de motifs et l'un des critères d'extraction. Le nombre de motifs extraits est toujours plus important en utilisant les valeurs-seuils des grilles DCE qu'avec celles du SEQ-eau. En revanche, il est nécessaire d'identifier les altérations stables – telle que la température de très bonne qualité ou en très bon état – ainsi que celles très fréquentes – telles que les phosphates de bonne qualité pour le SEQ-eau – et de les retirer avant extraction, afin d'obtenir des motifs discriminants ne se limitant pas à ces seules altérations, mais également pour ne pas dépasser les capacités de calcul de PRESTOR. En effet, sur l'ensemble de la France et de la période disponible, si la fréquence minimale est trop faible et la durée des successions temporelles trop longue, l'algorithme atteint ses limites et ne donne pas de résultats.

L'un des défis des méthodes de fouille est de hiérarchiser les résultats selon leur pertinence quant à la question posée. Nous proposons de combiner quatre mesures d'intérêts associées aux motifs : leur complexité, leur fréquence, leur singularité, leur émergence afin de sélectionner les motifs les moins triviaux, les plus discriminants et spécifiques de l'indice et de l'état biologique qu'ils précédent.

Nous avons exploré l'extraction réalisée sur l'ensemble de la France et de la période disponible, en utilisant le SEQ-eau, pour un intervalle de temps de 24 mois et une fréquence minimale de 0,6, pour deux indices biologiques : l'IBGN et l'IBMR. Les états biologiques des motifs sont, le plus souvent, pires que les qualités des altérations physico-chimiques qui les précédent, ce qui laisse supposer des synergies entre les différentes altérations présentes. Quels que soient les cinq indices biologiques considérés, les motifs précédant un très bon état biologique sont caractérisées par des altérations stables en très bon état (exemple: Figure 13). A l'opposé, il existe deux types des motifs précédant les mauvais états biologiques de l'IBGN et l'IBMR :

- des motifs aux altérations multiples et chroniques dont les nitrates, les pesticides, les micropolluants organiques hors pesticides de qualité moyenne, médiocre ou mauvaise, se répétant plusieurs fois (exemple : Figure 14) ;
- ou une seule apparition d'une altération dégradée comme les matières azotées hors nitrates de qualité moyenne, les matières organiques et oxydables de qualité médiocre, parmi d'autres altérations non dégradées (exemple:
- Figure 15).

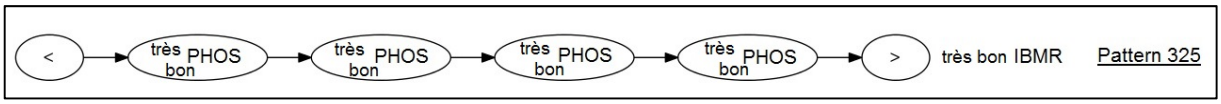


Figure 13 : Motif numéro 325 ayant précédé un IBMR en très bon état (critères d'extraction : toute la France, période 2007-2013, SEQ-eau, longueur de séquence 24 mois, fréquence minimale 0,6) ; altération PHOS : matières phosphorées

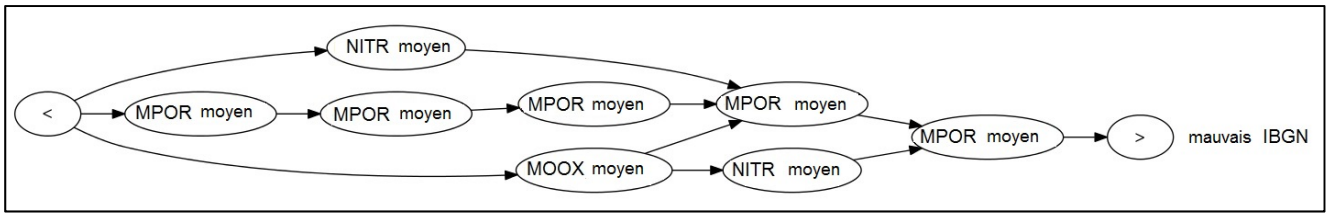


Figure 14 : Motif numéro 300 ayant précédé un IBGN en mauvais état (critères d'extraction : toute la France, période 2007-2013, SEQ-eau, longueur de séquence 24 mois, fréquence minimale 0,6) ; altérations MPOR : micropolluants organiques hors pesticides, MOOX : matières organiques et oxydables et NITR : nitrates

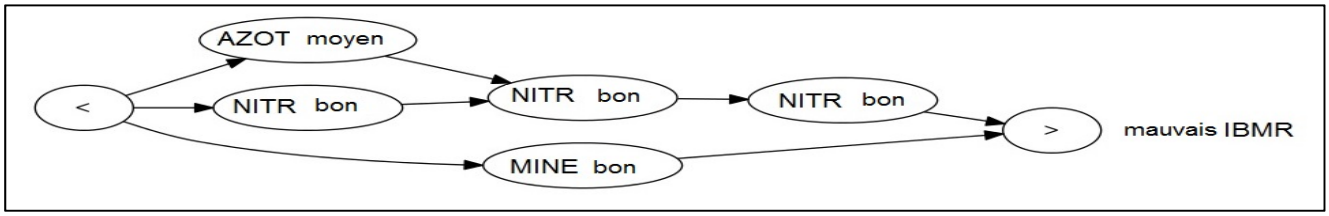


Figure 15 : Motif numéro 442 ayant précédé un IBMR en mauvais état (critères d'extraction : toute la France, période 2007-2013, SEQ-eau, longueur de séquence 24 mois, fréquence minimale 0,6) ; altérations AZOT : matières azotées hors nitrates et NITR : nitrates

L'altération micropolluants minéraux, c'est-à-dire les métaux lourds, est la seule altération qui n'est observée que dans les motifs de l'IBGN (exemple : Figure 16). Nous n'avons pas trouvé d'autre altération qui soit spécifique d'un indice biologique donné.

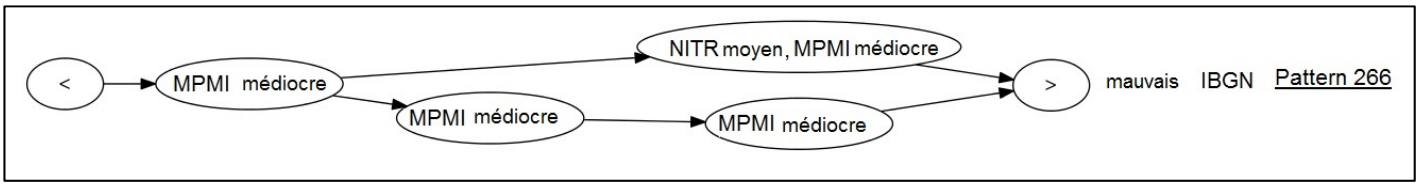


Figure 16 : Motif numéro 266 ayant précédé un IBGN en mauvais état (critères d'extraction : toute la France, période 2007-2013, SEQ-eau, longueur de séquence 24 mois, fréquence minimale 0,6) ; altérations MPMI : micropolluants minéraux et NITR : nitrates

Les motifs les plus discriminants sont obtenus pour les classes de qualité extrêmes, très bonnes et mauvaises, ce qui met en évidence la difficulté de fixer des valeurs-seuils de classe pertinente, que ce soit pour les indices biologiques ou pour les altérations. Nous ne présentons pas les résultats obtenus avec les grilles DCE, mais dans ces motifs, l'altération « nutriments » n'a jamais été autrement que dans le

bon état. Or, il est à noter que pour cette altération, la valeur-seuil du bon état est à présent de 50 mg/L pour les nitrates contre 10 mg/L dans le SEQ-eau.

Les motifs sont visuellement didactiques et apparaissent donc comme un outil intéressant pour identifier les séquences temporelles d'altérations qui ont conduit à un état biologique donné dans des contextes de pressions multiples, un des challenges de l'application de la DCE.

2 Article

TITLE: **Mining the sequential patterns of water quality preceding the biological status of waterbodies**

Authors:

Corinne Grac[a,b], Xavier Dolques , Agnès Braud[c], Michèle Trémolières[b], Jean-Nicolas Beisel[a,b], Florence Le Ber[c,d]

a ENGEES, LIVE UMR 7362, F-67000 Strasbourg, France
corinne.grac@engees.unistra.fr
b Université de Strasbourg, CNRS, LIVE UMR 7362, F-67000 Strasbourg, France
c Université de Strasbourg, ICube UMR 7357, F-67400 Illkirch-Graffenstaden, France
d ENGEES, ICube, UMR 7357, F-67000 Strasbourg, France

2.1 Abstract

We implemented a specific data mining process to explore the relationship between biological indices and physico-chemical pressures in rivers. Data were collected in the framework of the French National monitoring network set up to assess the ecological status of rivers under the European Water Framework Directive (WFD). Chemical parameters and biological indices were collected regularly from 1.781 locations in metropolitan France from 2007 to 2013. The sequential pattern mining process generates closed partially ordered patterns representing a succession of physico-chemical events that precede a given biological index in a given status, validated using a subset of data. This paper focuses on the patterns and their occurrence. The three main advantages of this tool are: (i) its rapidity and efficiency, (ii) its ability to highlight relationships between the physico-chemical and biological states which occur over time, (iii) its specific ability to highlight a sequence of alteration events before an observed biological response. Thus the physico-chemical statuses of water bodies usually appeared to be higher than their biological statuses, suggesting synergism between toxicants and/or an additive impact of other stressors related to hydromorphology or hydrology. Patterns found in the highest

biological status for the biological indices based on macroinvertebrates, diatoms, macrophytes or fish, were characterised by the constancy of a high physico-chemical status over time. By contrast, before indices based on macroinvertebrates and macrophytes, two types of patterns were observed for bad biological status: (1) a chronic multi-pressure pattern, in which pressure categories such as nitrates, pesticides and other organic hydrocarbons, in moderate, poor or bad status, repeated themselves several times over time, or (2) a single occurrence of a degraded pressure category, such as one moderate nitrogen, excluding nitrate, or one poor oxidizable organic matter, among other pressure categories in good status. The patterns we obtained are a promising solution both to disentangle the effects of the different stressors on water quality, and to identify the key temporal sequences among them in a context of multi-stress conditions, which is a challenge currently facing the WFD. But some values of thresholds used to discretize physico-chemical or biological status could be discussed.

2.2 Keywords

Rivers, water quality, biological status, data mining, temporal patterns, physico-chemical status

2.3 Introduction

In rivers, several human activities produce a combination of pressures which alter the abiotic components of the ecosystem, affect the biological communities and hence their ecological status (driving forces, pressure, state, impact, response, i.e. the DPSIR concept, Kristensen, 2004). Multi-stress conditions are multiscale (Dahm et al., 2013; Lalande, 2013). In Europe, the challenge currently facing the European Water Framework Directive (WFD) (European Council, 2000), which requires the achievement of a good ecological status for the conservation or restoration of aquatic ecosystems, is to have access to biological tools that not only able to assess this status but also to disentangle and identify the different pressures and to be able to propose the appropriate restoration actions to achieve good ecological status (Reyjol et al., 2014; Feld et al., 2016). Furthermore, the assessment of aquatic ecosystems relies on monitoring, which generates large volumes of heterogeneous data from multiple sources (Hering et al., 2010) at different temporal scales. Data mining methods are able to analyse large datasets and may be a good alternative to traditional statistical methods (Giraudel and Lek, 2001). These methods can produce readable results, thereby facilitating interactions between data miners and experts (Džeroski et al., 1997). Two categories of data mining methods exist: supervised and unsupervised methods. Supervised methods require a learning dataset to build a specific model adapted to a given issue, including variables and expected results. When the model is built, it has to be tested on a second dataset. After this validation stage, it can be applied to a third dataset to predict results expected in similar conditions. Unsupervised methods do not require a learning dataset and make it possible to explore and identify unexpected rules in a dataset.

Several authors (Larras et al., 2017; Mondy and Usseglio-Polatera, 2013; Villeneuve et al., 2015) have already used water quality data to test the potential ability of supervised data mining methods to identify key anthropic pressures using different biological organisms. The search for predictive correlations using supervised learning requires the identification of explanatory and target variables among the descriptors originating from the process described above. Once this step is completed, a large set of supervised learning techniques is available, including some artificial neural networks (D'heygere et al., 2006; Dakou et al., 2006; Everaert et al.,

2016; Tsai et al., 2016) and decision trees (Džeroski, 2001; Feld et al., 2016a; Larras et al., 2017; Mondy and Usseglio-Polatera, 2013; Villeneuve et al., 2015). However, these off-the-shelf tools still need to be adapted to the nature of the available data as well as to the specific problem at hand. Furthermore, spatial and temporal dependencies require dedicated methods including Markov random fields or Bayesian networks (Adriaenssens et al., 2004; Forio et al., 2016b; Fytilis and Rizzo, 2013; Landuyt et al., 2016a; Van Looy et al., 2015). State-of-the-art reviews underline the importance of considering and combining biological and physico-chemical variables in order to discover relevant knowledge (Marzin et al., 2012; Oberdorff and Hughes, 1992). Yet none of the studies cited above accounted for temporal aspects in the way temporal pattern mining approaches do, which is the best way to analyse disturbances and biological compartment dynamics.

To explore data resulting from monitoring the quality status of rivers, we propose an unsupervised data mining tool able to treat large volumes of data and to account for the temporal aspect of events: sequential pattern mining. The tool was first introduced by Agrawal and Srikant (1995) and is a temporal extension of association rules (Agrawal and Srikant, 1994), which were first developed for and used in engineering software (Ren et al., 2009), medicine (Sallaberry et al., 2011), and marketing (George and Binu, 2012). In order to reduce information redundancy in sequential patterns and to limit their exponential numbers and hence the volume of the result, Fabrègue et al. (2013) proposed an adaptation: closed partially ordered patterns (CPO-patterns). CPO-patterns can be used in all kinds of sequential databases and have three main advantages: (1) they provide more detailed information on order among elements; (2) they are depicted by a directed acyclic graph, which is easy to understand; (3) they summarise sequential pattern sets.

In the present study, we focus on an implementation of a tool generating CPO-patterns with the aim of exploring the relationship between biological indices in different statuses and the succession of physico-chemical events which precede them in order to disentangle and identify the pressure categories implicated in the

case of not good status. Based on the results obtained with the use of patterns, we discuss the contributions, the limits and the potentials of this original approach.

2.4 Materials and methods

2.4.1 The dataset used

Data were collected in 1,781 sampling sites (Figure 17) in the framework of the French network created to assess the ecological status of waterbodies according to the WFD. The WFD requires the achievement of a good (ecological and chemical) status for the conservation or restoration of waterbodies (management unit of the WFD) in the short (2021) and medium term (2027). Referring to running waters, the WFD defines a waterbody as "a discrete and significant element of surface water such as […] a stream, river or canal, part of a stream, river or canal […]". The WFD defines ecological status as an expression of the quality of the structure and functioning of inland aquatic ecosystem (WFD, art. 2).

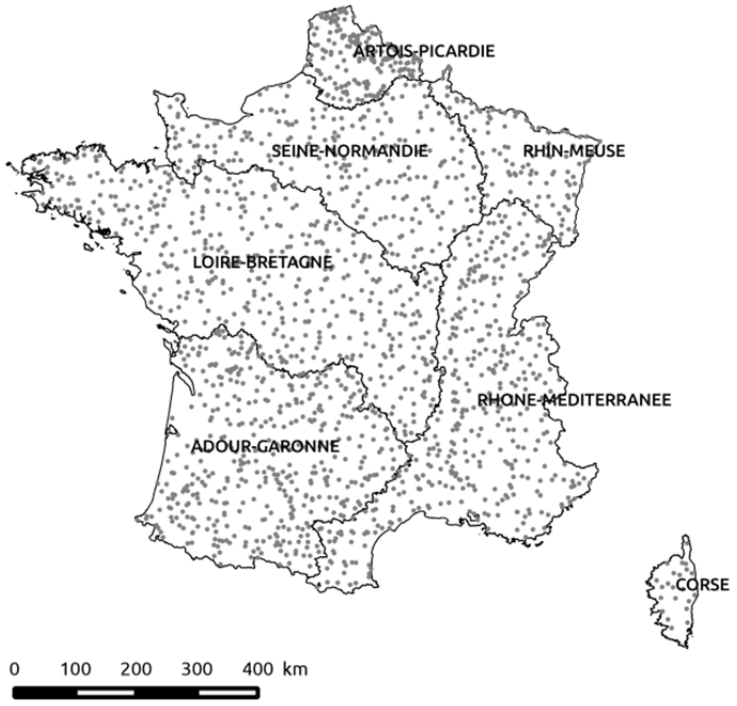


Figure 17: Location of the 1,781 French sampling sites for the national ecological assessment of rivers

Fieldwork was undertaken by regional environmental agencies on the French national river network between 2007 and 2013. Table 4 lists the volume of the principle data we used. The 23,071,909 physico-chemical results concerned 1,201 parameters. Among them, 2% were major parameters (e.g. pH, nitrogen) and 98% were minor parameters, including micro-pollutants (e.g. copper, atrazine). Analyses of the major elements were conducted 12 times a year and analyses of minor elements four or six times a year; but the completeness of the physico-chemical data varied consideraly over the study period, ranging from 100% to 1% (1% of parameters were than 80% complete, 57% between 80% and 10%, 26% between 10% and 1%, 16% were less than 1% complete). Above, we used the physico-chemical results which were the most complete: they represented 1,146,544 results and concerned 189 parameters (completeness ranged between 98% and 33% for polychloro-biphenyls, PCB). The 24,593 biological results concerned four biological groups and five French standardised biological indices (identified below by their French acronym): macroinvertebrates (33% of results) with IBGN (AFNOR, 2004a) and I2M2 (AFNOR, 2016, 2010), diatoms (35%) with IBD (AFNOR, 2007), fishes (23%) with IPR (AFNOR, 2004b), macrophytes (10%) with IBMR, (AFNOR, 2003). Calculation of I2M2 is detailed in Mondy et al. (2012). IBGN, the former French macroinvertebrate index is expected be replaced by I2M2, but at present, the two are still calculated. The frequency of biological sampling was once a year for macroinvertebrates and diatoms, once every two years for fishes and macrophytes. The overall biological results were 69% complete.

Table 4: Volume of data used

Type of data	Number
Sampling site	1,781
Water sampling for physical chemical analyses	122,765
Biological sampling	26,072
Physical chemical parameters	1,201
Physical chemical results	23,071,909
Biological groups	5
Biological index results	24,593

We created a database specifically for these data.

2.4.2 Data pre-treatment

We discretized chemical and biological data using the five levels represented by colors, which symbolize the different statuses of the WFD; i.e. blue: high, green: good, yellow: moderate, orange: poor, and red: bad.

For the physico-chemical data, we used two sets of thresholds: the first was created in France before the application of the WFD, hereafter referred to as SEQ (MEDD and AE, 2003) and a second one upon application of the WFD hereafter called the WFD guide (MEEM, 2012). We performed physico-chemical discretization following a three-step process for a given date: (1) by grouping parameters in the pressure categories listed in Table 5 for SEQ and in Table 6 for the WFD guide; (2) by discretization of each parameter in each pressure category according to the corresponding thresholds; (3) by attributing the final level of each category keeping the worst level of parameters considered, based on the one out, all out principle. With SEQ, we used 192 physico-chemical parameters and with the WFD guide, 58 parameters. The thresholds of the two sets are listed in appendix 1 (for SEQ) and appendix 2 (for the WFD guide).

Table 5: List of physico-chemical pressure categories, their acronyms based on SEQ (MEDD and AE, 2003) and the number of associated parameters

N°	Acronym	Pressure categories	N° of parameters
1	MOOX	Oxidizable organic matter (e.g. O_2, DBO)	7
2	AZOT	Nitrogen excluding nitrate	3
3	NITR	Nitrate	1
4	PHOS	Phosphorous	2
5	EPRV	Effect of eutrophication	2
6	PAES	Suspended matter	2
7	TEMP	Temperature	1
8	ACID	Acidification parameters	2
9	MINE	Mineralisation	8
10	MPMI	Heavy metals	10
11	PEST	Pesticides	74
12	HAP	Polycyclic aromatic hydrocarbons	15
13	PCB	Polychloro-biphenyls	8
14	MPOR	Other organic hydrocarbons	57

Table 6 : List of physico-chemical pressure categories, their acronyms based on the WFD guide (MEEM, 2012) and the number of associated parameters

N°	Acronym	Pressure categories	N° of parameters
1	TEMP	Temperature	1
2	ACID	Acidification parameters	1
3	BILO2	Oxygen balance	4
4	NUTRI	Nutrients	5
5	POSPE	Specific pollutants	9
6	SDP	Priority substances and priority hazardous substances	38

For each biological index, we took the thresholds from MEEM (2012). Figure 18 shows the distribution of the available data according to their status in each index. In most cases, about two-thirds of the results correspond to the status required by the WFD guide (class 1, high status, or class 2, good status) while one third of the results are not good quality (Class 3, moderate status; class 4: poor status; class 5: bad status) except for IPR (52% in classes 1 and 2 and 48% in classes 3, 4 and 5).

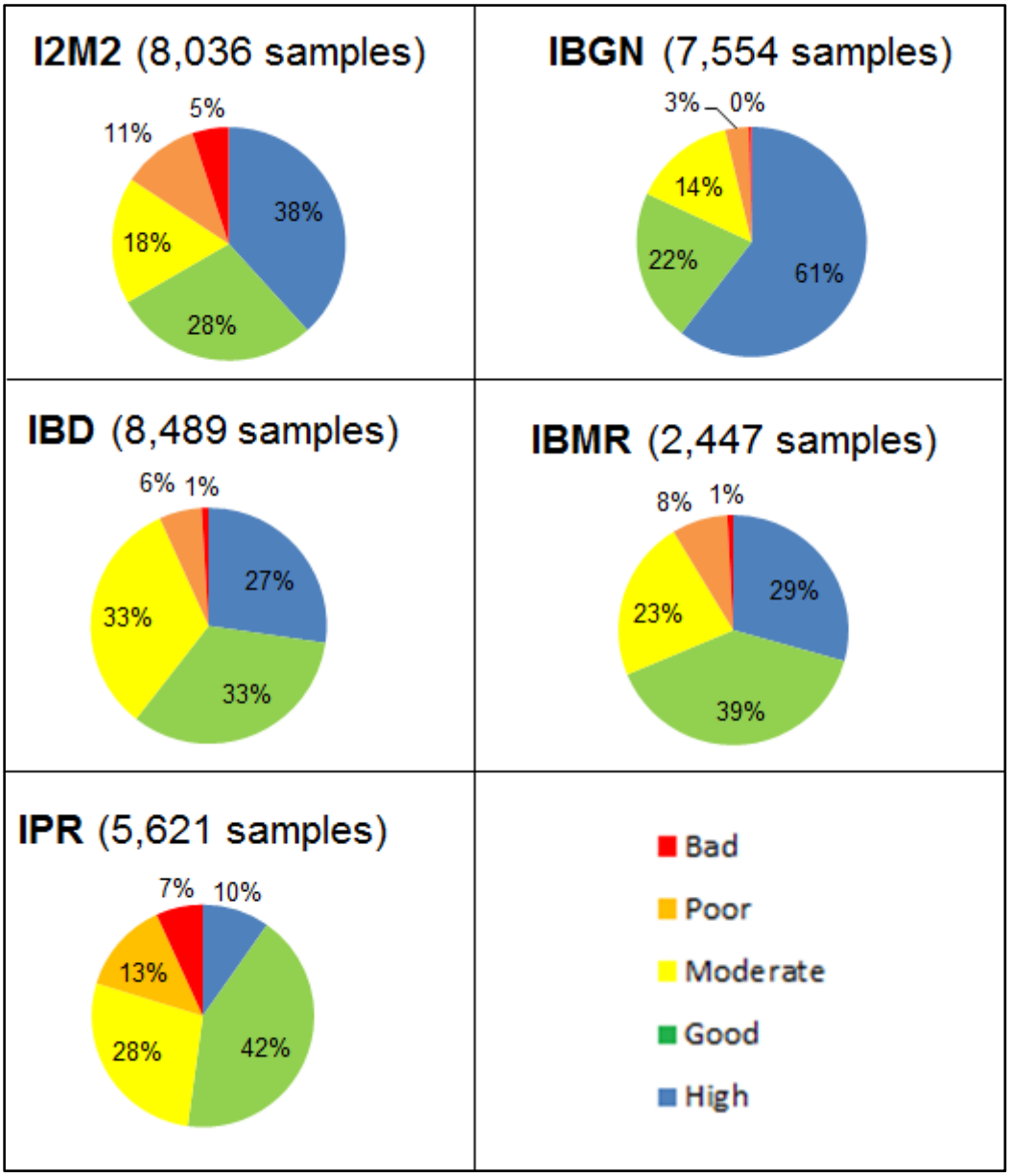


Figure 18: Distribution of status classes according to French biological indices (macroinvertebrate index: I2M2 and IBGN, diatom index: IBD, macrophyte index: IBMR, and fish index: IPR) in the national dataset used

2.4.3 The pattern extracting tool: "PRESTOR"

PRESTOR **(temporal PRESsure categories patterns extracTOR)** is a tool for the analysis of a huge database, developed in C++ for the best performances (along with intel TBB for parallel programming and BOOST and QT to handle different kinds of interfaces). PRESTOR is a command line one, most of the configuration is done in a "config.ini" file. It can fetch data directly from the database or from well-formatted files. Several configuration options are provided (described below) to select specific data subsets. The result consists in pictures and a few files providing different kinds of statistical values to help understand the results.

Frequent Closed Partially Ordered Patterns

The main algorithm at the core of PRESTOR generates frequent closed partially ordered patterns (CPO-patterns) as proposed by Fabrègue et al. (2013), which is an improvement on the sequential patterns created by Agrawal and Srikant (1995).

A sequential pattern is a succession of physico-chemical events that can be checked by the samples preceding a biological status assessment at a sampling site on one or more occasions. Formally, it is a sequence of itemsets. An item is a value to be recognised and an itemset is a set of values assessed at the same point in time (values from the same sample).

For example, having at the same time a *high* value for the alteration *PHOS* (noted *High PHOS*) and a bad value for the alteration *MPOR* (noted *Bad MPOR*) is an itemset we denote (*High PHOS*, *Bad MPOR*). <(*High TEMP*),(*High PHOS*, *Bad MPOR*)> is then a sequential pattern checked at all stations where the itemset (*High TEMP*) is found before the itemset (*High PHOS*, *Bad MPOR*).

Table 7: Invented database presenting the history of 3 sampling sites

Station	Date	Value
1	2/6/2015	*Bad PHOS*
1	31/5/2015	*High IBGN*
1	29/5/2015	*High PHOS*
1	29/5/2015	*Bad MPOR*
1	2805/2015	*Moderate AZOT*
1	27/5/2015	*Poor MINE*
1	27/5/2015	*High TEMP*
2	28/6/2015	*High IBGN*
2	27/6/2015	*High TEMP*
2	27/6/2015	*High PHOS*
2	27/6/2015	*Bad MPOR*
2	26/6/2015	*Moderate AZOT*
3	30/4/2015	*High IBGN*
3	28/4/2015	*Bad MPOR*
3	27/4/2015	*High PHOS*
3	27/4/2015	*High TEMP*
3	26/4/2015	*Moderate AZOT*
3	18/3/2014	*Good TEMP*

Let us consider the preceding (invented) database presenting the history of 3 stations (Table 7). In this paper, we focus on the biological indices (e.g. IBGN) we want to characterise using the preceding samples. To this end, we build sequences of itemsets from this database following two simple rules: the sequence must end on a biological sample, called the generating context, and the duration in months between the first and the last samples is bounded by a value (here 6 months) defined by the operator. Three sequences were extracted:

-S1: <(*High TEMP*, *Poor MINE*),(*Moderate AZOT*),(*Bad MPOR*, *High PHOS*),(*High IBGN*)>

-S2: <(*Moderate AZOT*),(*Bad MPOR*, *High PHOS*, *High TEMP*),(*High IBGN*)>

-S3: <(*Moderate AZOT*),(*High TEMP*, *High PHOS*),(*Bad MPOR*),(*High IBGN*)>

The sample collected at station 1 on 2/6/2015 was not included in a sequence as it did not precede a biological sample. The sample collected at station 3 on 18/3/2014 was not included in a sequence even though it did precede a biological sample because it was too old with respect to the time-length threshold.

In the sequence database, only sequence S1 confirmed the sequential pattern <(*High TEMP*),(*High PHOS*, *Bad MPOR*)> (it was said that {S1} supported the pattern), even though other itemsets exist between (*High TEMP*) and (*High PHOS*, *Bad MPOR*).

Here we consider only the order of events, this is why sequences S2 and S3 did not confirm the sequential pattern. On S2 *High TEMP* occurred at the same time as *High PHOS* and *Bad MPOR* whereas to confirm the pattern, it would need to occur before. On S3, the problem was that *High PHOS* and *Bad MPOR* did not occur at the same time.

For a set of sequences, several sequential patterns may be valid e.g. <(*Moderate AZOT*)(*High PHOS*),(*High IBGN*)>; <(*Moderate AZOT*)(*Bad MPOR*),(*High IBGN*)>; <(*High TEMP*),(*High IBGN*)>; <(*High PHOS*),(*High IBGN*)>; <(*Moderate AZOT*),(*High IBGN*)>; <(*Bad MPOR*),(*High IBGN*)>; <(*High IBGN*)>; <(*Moderate AZOT*)(*High PHOS*)>; <(*Moderate AZOT*)(*Bad MPOR*)>; <(*High TEMP*)>; <(*High PHOS*)>; <(*Moderate AZOT*)>; <(*Bad MPOR*)>; <> were all valid sequential patterns confirmed by sequences S1, S2 and S3.

However, some patterns may be redundant as the information they convey is already contained in other patterns. For example, the information in <(*High PHOS*),(*High IBGN*)> is already conveyed by the pattern <(*Moderate AZOT*)(*High PHOS*),(*High IBGN*)>. Hence, there is always a smallest set of sequential patterns conveying all the information regarding a set of sequences. A sequential pattern from this set is called a closed sequential pattern. A sequential pattern is closed with respect to a set of sequences if no other pattern conveys the same information, e.g. {<(*High TEMP*),(*High IBGN*)>,<(*Moderate AZOT*)(*High PHOS*),(*High IBGN*)>,<(*Moderate AZOT*)(*Bad MPOR*),(*High IBGN*)>} is the set of closed sequential patterns for sequences S1, S2 and S3; which means that it contains the same information as the set of all sequential patterns presented earlier.

Closed sequential pattern sets are relevant because they represent the smallest and most complete description possible of a set of sequences with sequential patterns.

In order to facilitate the reading of closed sequential pattern sets, Fabrègue et al. (2013) summarised them in a single structure named closed partially ordered patterns (CPO-patterns). A CPO-pattern is a directed acyclic graph in which each path is a sequential pattern. It is built in such a way that the number of nodes is minimised.

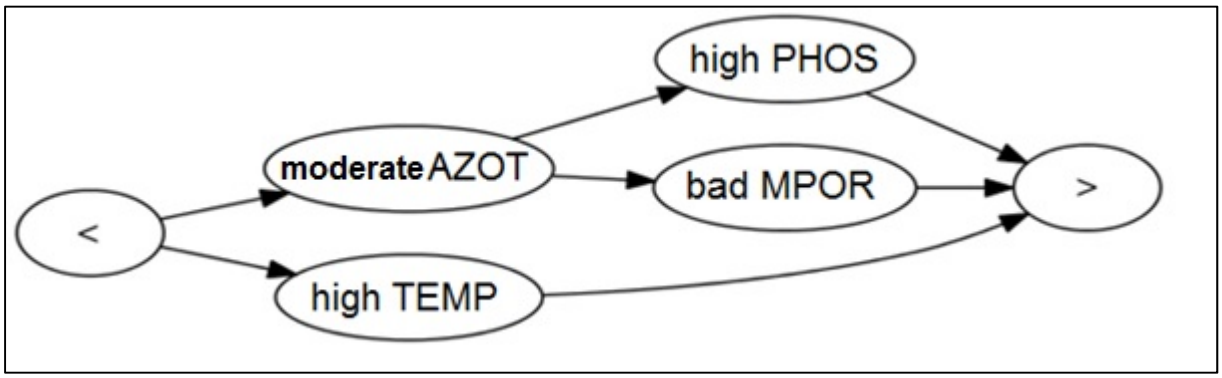


Figure 19 : CPO-pattern for sequences S1, S2 and S3

The algorithm extracts the CPO-patterns (Figure 19) only for a specific biological index in a specific status, here *High IBGN,* which is the generating context. Hence we did not display the generating context of the pattern but this is important information for the evaluation. The CPO-pattern obtained is interpreted as follows: for each sequence which confirms it, *High IBGN* is preceded by *High TEMP* and by *Moderate AZOT* followed by *High PHOS* and *Bad MPOR* in any order. All the paths in the graph are checked by referring to all the supporting sequences.

Implementing the method proposed by Fabrègue et al. (2013), PRESTOR extracts all the frequent CPO-patterns for a given sequence database. The frequency of a pattern is the ratio of the size of its support to the size of the sequence database. In some extracts, several CPO-patterns extracted in several contexts may be identical. Each has a generating context, e.g. *High IBGN*, *High IBMR*. For this set of identical CPO-patterns, the context of the CPO-pattern with the highest frequency is called the dominant context.

Hereafter, we refer to CPO-patterns simply as patterns.

Choice of parameters (region, time, reference thresholds, etc.)

The operator can modify several parameters to obtain different kinds of results. A filter can be applied to the data to limit the analysis to a subset of the database restricted to a given region or a period of time. It is also possible to use different quality reference norms for discretizing the physico-chemical values (SEQ or the WFD guide). The operator has to define a frequency threshold below which the patterns are not extracted. The algorithm itself can also be parameterized to consider different time lengths (in months) for the input sequences. One can also limit the values considered to remove irrelevant values. One example of an irrelevant value is *High TEMP* because it appears in more than 90% of the sequences in the database. This generates a significant number of parasite nodes which provide no information. Note that only the *High TEMP* is filtered but *Bad TEMP* is kept as such because this lowest general value may provide some information. The choices of these parameters for an extraction is its configuration.

2.4.4 Metrics for the evaluation of patterns

To generate a batch of patterns with PRESTOR, we need to (1) delimit the dataset according to its geographical dimension (e.g. all France) and the time period to be considered (e.g. to 2007 to 2013), (2) specify the quality reference for discretizing the physico-chemical value (e.g. SEQ), (3) specify the time-length before the targeted biological value (e.g. 24 months) and the minimum frequency expected (e.g. 0.6). PRESTOR automatically creates a folder which contains all the results produced: (1) the file "Configuration", (2) n picture files each representing a pattern (e.g. 788 picture files), (3) a file "Contexts" which gives, for each context, the count of sequences found and sampling sites found (e.g. high I2M2: number of sequences = 3,056 and number of sampling sites: 805), (4) a table "Results" giving the frequency of each pattern in each context, (5) a table "Supports" giving the list of all sampling sites at each date concerned by each pattern (Figure 20).

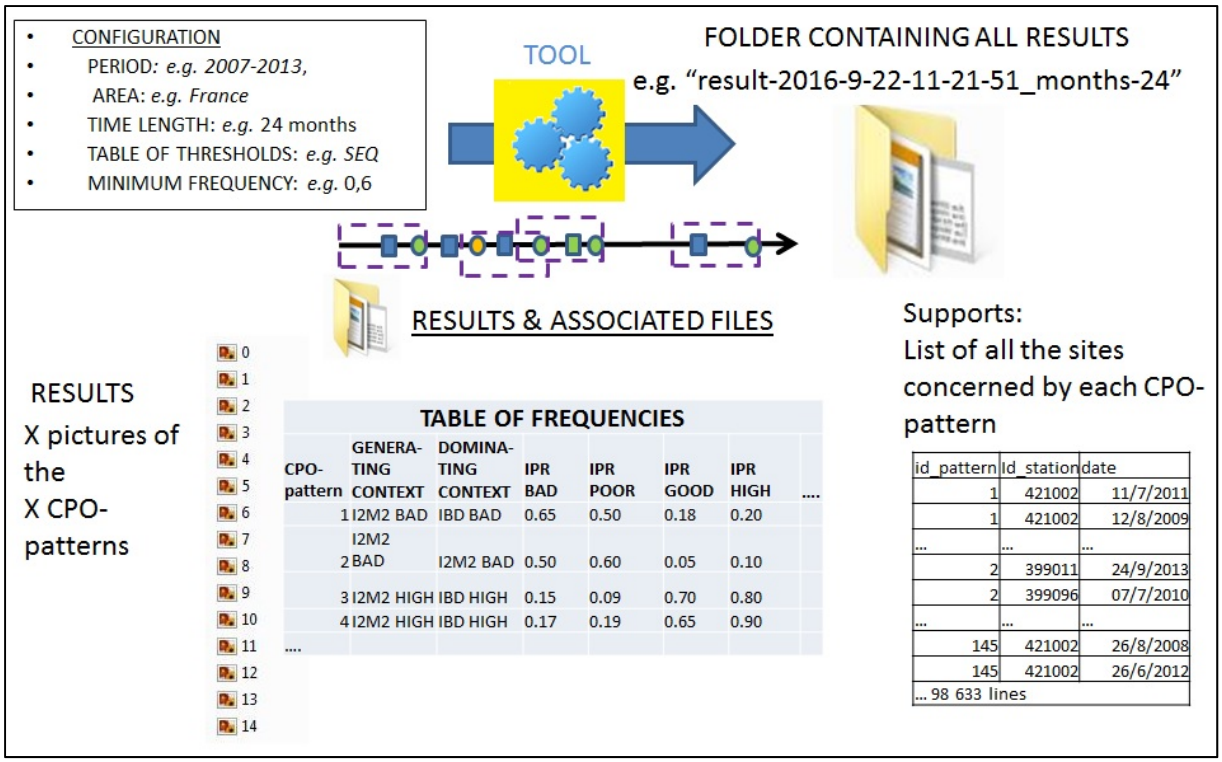


Figure 20: Diagram of the results and the associated files generated by PRESTOR when extracting patterns

In addition to the frequency (f), we propose three metrics of interest per pattern to describe the results: 1) their complexity (C) (Equation 1), 2) their scarcity (S) (Equation 2), 3) their emergence (E) (**Equation 3**).

The complexity (C) of a pattern P indicates its relative size compared to the biggest pattern found in a given extraction. Its value range is [0-1]: it is zero when the pattern P is empty and 1 when pattern P is the biggest one.

Equation 1 $$C = n / N$$

where n is the number of items in pattern P; N is maximum number of items found in the biggest pattern in the chosen extraction.

Scarcity (S) conveys the level of specificity of pattern P in its generating context. Its value range is [0-1]: it is zero when pattern P is found in all 25 contexts and 1 when pattern P is found only in one context.

Equation 2 $S= (1-n'/25)$

where n' is number of repetitions of the pattern in other extraction contexts.

The emergence (E) of pattern P, in a given extraction, is calculated only if the generating context is dominant ($f=f_{max}$). The bigger E is, the more specific P is for its targeted biological value. Its value range is [0-∞].

Equation 3 $E= f_{max}/f_{(max-1)}$

2.5 Results

2.5.1 Overall performance of PRESTOR

We generated 117 pattern extractions with different configuration parameters and obtained a total of 52,374 patterns. In our conditions (Computer INTEL COR I7-4790 3.6 Go and 16 Go RAM), extractions take from few seconds to few minutes, depending on the size of the dataset and the configuration parameters. Only 7% of these extractions failed due to too low minimum frequency. They succeeded with a higher minimum frequency. This low minimum frequency changed with the configuration parameters: it was lower than 0.4 for the following configuration [Area France, period 2007-2013, time-length 6 months, table of thresholds: SEQ] and lower than 0.6 for the same configuration except when the table of thresholds came from the WFD guide. The results were considered stable when a new extraction with a given configuration produced the same results.

As shown in Figure 21, the number of patterns increases with an increase in time-length and with a decrease in the minimum frequency or in the number of sampling sites, but without constant correlations. The number of patterns obtained using the WFD thresholds guide is always higher than with SEQ thresholds in the same configuration (see config-5 and config-6 on graph C, Figure 21).

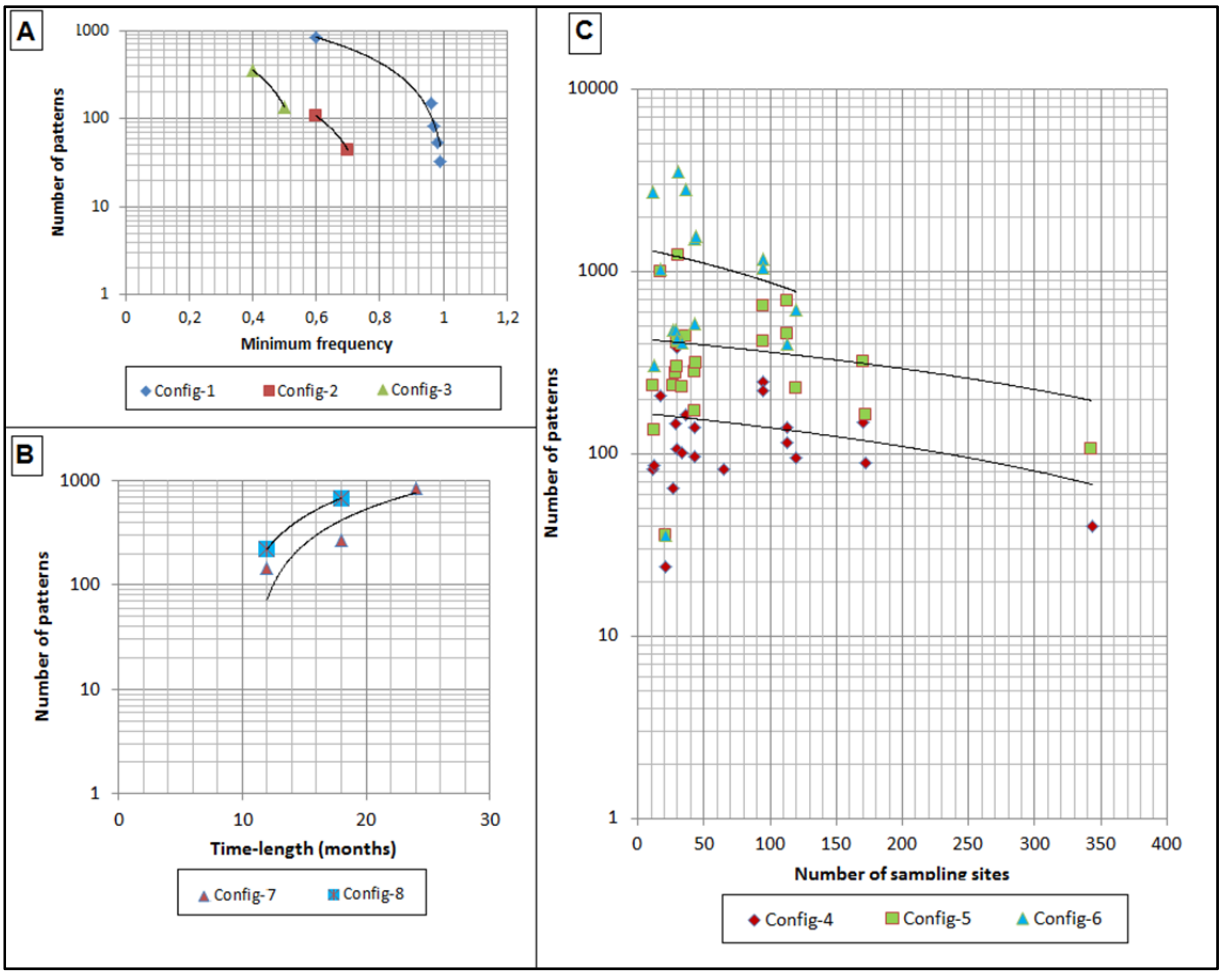


Figure 21: Variation in the number of patterns according to: A: variation in minimum frequency, B: variation in time-length, C: variation in the number of sampling sites; extractions obtained with 8 configurations for period: 2007-2013: Config-1& 3 = [area: France, table of thresholds: SEQ, min. frequency: VARIABLE, Config-1, time-length: 24 months, Config-3, time-length: 6 months]; Config-2= [area: Paris and surroundings, table of thresholds: SEQ, time-length: 6 months, min. frequency: VARIABLE]; Config-4 & 5 = [area: VARIABLE, table of thresholds: SEQ, time-length: 12 months, Config-4 min. frequency: 0.8; Config-5 min. frequency: 0.6]; Config-6 = [area: VARIABLE, table of thresholds: WFD guide, min. frequency: 0.6]; Config-7 & 8= [area: France, time-length: VARIABLE, Config-7 table of thresholds: SEQ min. frequency: 0.6], Config-8 table of thresholds: WFD guide, min. frequency: 0.8]

2.5.2 *Batch of patterns in a specific configuration*

Here we present the results obtained with the following configuration: [area: France, period: 2007-2013, table of thresholds: SEQ, time-length: 24 months, minimum frequency: 0.6]. We first eliminated 4 pressure categories for which a single status dominates over 80%: TEMP, ACID and EPRV in high status and HAP in moderate status to avoid patterns with repetitive and no discriminating items.

A first batch of patterns was generated for a minimal frequency of 0.96. As 90% of the patterns obtained contained only five pressure categories in a single status, high or good (good PAES, high MINE, good MOOX, good AZOT and good PHOS), hence masking the impact of other parameters, we also eliminated these pressure categories in these statuses from the input data. A second batch of patterns was generated without these pressure categories, for a minimum frequency of 0.6.

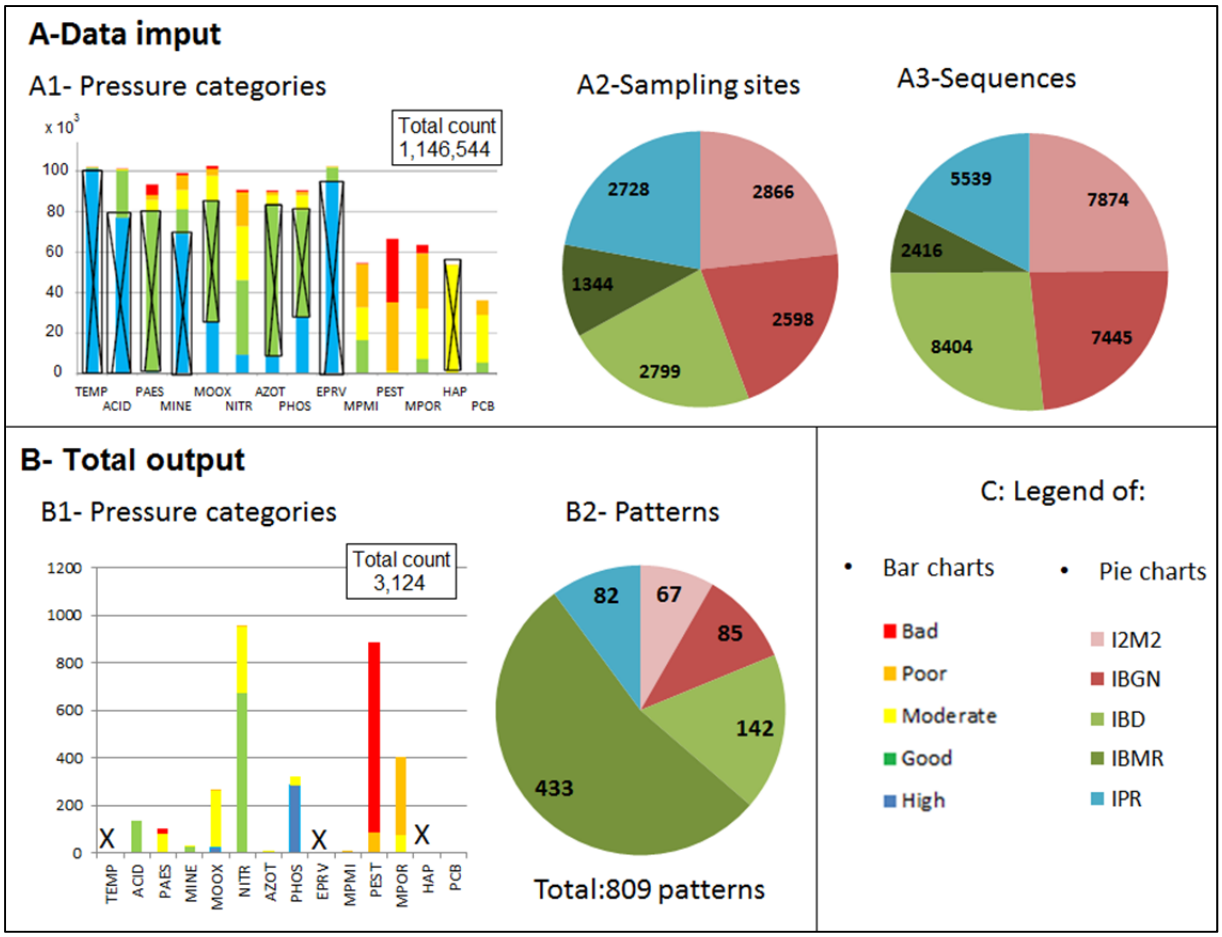


Figure 22: Data input (A) and total output (B) obtained with the configuration: [area: France, period: 2007-2013, table of thresholds: SEQ, time-length: 24 months, minimum frequency: 0.6]; A1- Number of results available per pressure category in the input dataset and distributed according to the status class, *crossed pressure categories removed before the patterns were extracted*; A2: Number of sampling sites per biological index (5 French indices: I2M2, IBGN, IBD, IBMR & IPR) available in the input dataset (a site being counted as many times as it changes status in an index); A3: Number of sequences available per biological index in the input dataset; B1: Number of pressure categories obtained in patterns and distributed per status class; B2: Number of patterns distributed per biological index.

The distribution of the alteration categories in the different statuses, in the patterns differed from those in the input data (Figure 22). The number of pressure categories in the input data was the same (8-9%: from 90,681 results for NITR to 102,006 results for TEMP compared to the total number of physico-chemical results: 1,146,544), while nitrogen was dominant in the patterns (30% = 953/3,124). The micro-pollutants were less abundant in the input data (from 3% -count: 36,158- for PCB to 6% -count: 66,523- for PEST). Conversely PEST was abundant in the patterns (29%). The distribution of sequences per index varied from 2,416 (8%) for IBMR to 8,404 (26%) for IBD, while 67 (8% of) patterns had I2M2 as context and 433 (i.e. 53%) had IBMR as context.

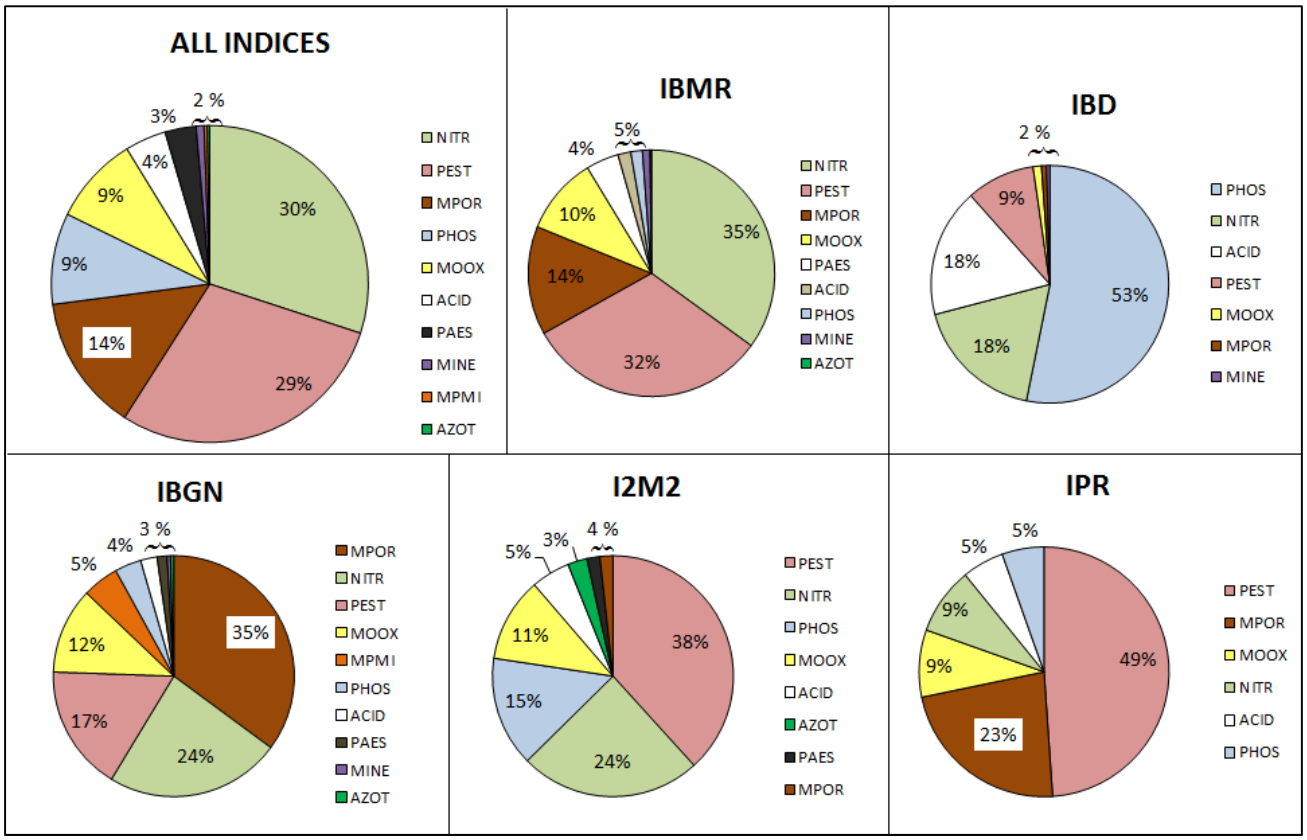


Figure 23: Overall distribution of the pressure categories in the 809 patterns obtained for the configuration [area: France, period: 2007-2013, table of thresholds: SEQ, time-length: 24 months, minimum frequency: 0.6] according to indices and ranked according to the top down percentage of appearance in patterns.

Five main pressure categories were found in the 809 patterns (Figure 23 – All indices): NITR (the most numerous first pressure category according to the top down distribution with 30% in all patterns) only in good and moderate statuses, PEST (the second one with 29% in all patterns) only in poor and bad statuses, PHOS and MOOX (with 9.0 % in all patterns) mainly in the moderate status – it will be recalled that MOOX with good status was removed before the extraction because of its dominance. We found the same distribution of four pressure categories for IBMR, with PHOS disappearing and PAES coming in fourth with 4%). Two main pressure categories were found in the IBD index: PHOS (with 53%, and fourth with 9.3% in all patterns) and ACID (sixth with 4% in all patterns). By studying patterns on the IBD index in detail, we observed that PHOS was mainly in high status - PHOS in good

status was removed before the extraction because of its dominance - and ACID (sixth with 4% in all patterns) only in good status –ACID in high status was removed before the extraction because of its dominance. MPOR (third with 14% in all patterns) was found in moderate status mainly in the IBGN index and in poor status in the IBMR and IPR indices. MPMI with 35% was found only in the IBGN index in poor status while PEST was the main pressure which impacts I2M2, the second index based on macroinvertebrates as IBGN. AZOT was found only in the I2M2 index in moderate status. The first pressure categories NITR, PEST, MPOR and MOOX representing at least 45% of the total distribution were the same in the IBMR, IPR and IBGN index but not in the same order of importance. PHOS, NITR and ACID were the main pressures in IBD, PEST, NITR and PHOS in I2M2.

2.5.3 *Characteristic patterns of IBGN and IBMR indices*

A total of 85 characteristic patterns were obtained for IBGN (Figure 24), (respectively 10, 6, 10, 18 and 46 in biological statuses high, good, moderate, poor and bad) and 433 patterns for IBMR (respectively 16, 12, 14, 88 and 303 in biological statuses high, good, moderate, poor and bad).

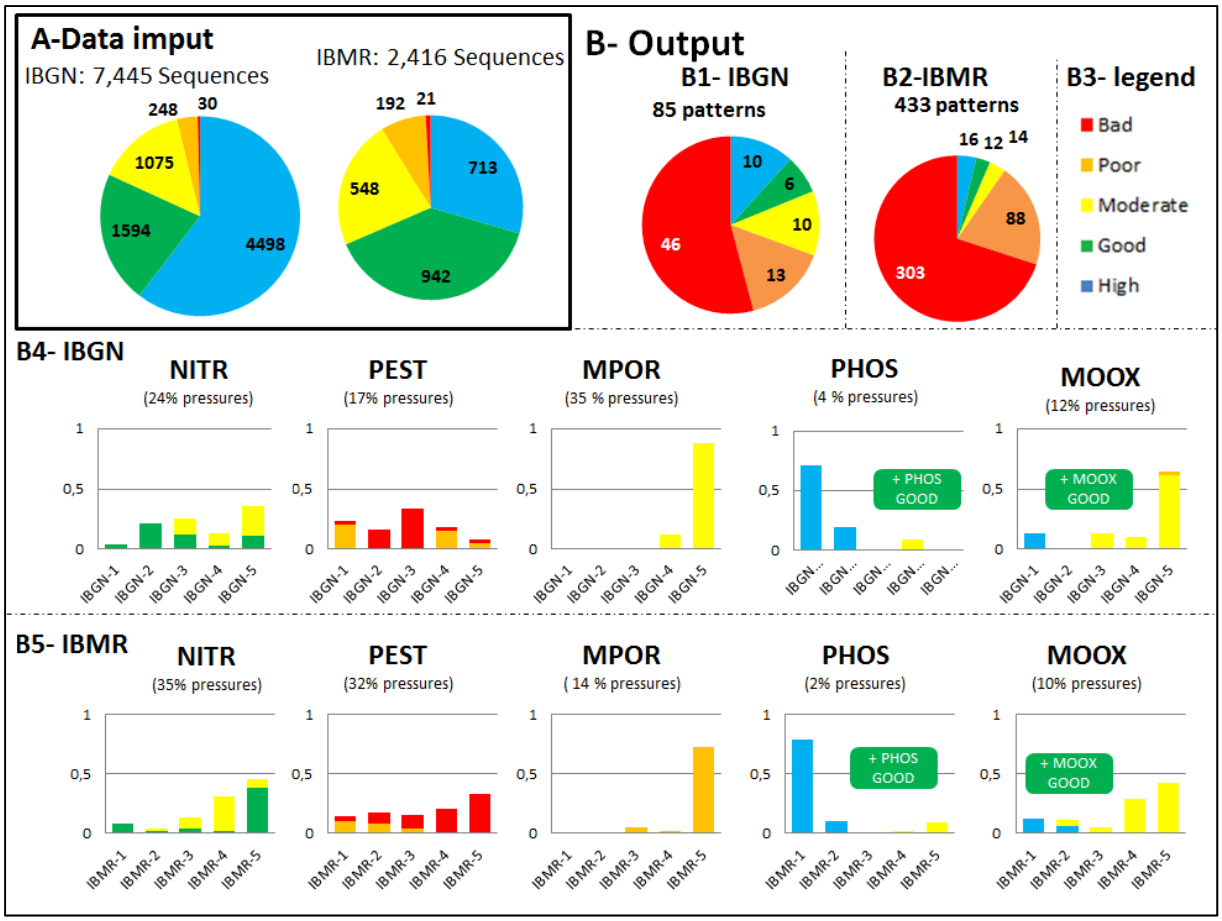


Figure 24: Data input (**A**) and total output (**B**) obtained with the configuration: [area: France, period: 2007-2013, table of thresholds: SEQ, time-length: 24 months, minimum frequency: 0.6] for the two French biological indices IBGN and IBMR; **A**- Number of sequences available for IBGN (left) and IBMR (right) in the input dataset; **B1 and B2**: Number of patterns obtained for IBGN (left) and IBMR (right) distributed per status class, respectively 85 and 433 patterns overall; **B4, B5**: distribution (**y'**) of the 5 status classes in the 5 first pressure categories in the global distribution (NITR, PEST, MPOR, PHOS and MOOX) in each status; B4 IBGN (IBGN-1: high, IBGN-2: good, IBGN-3: moderate, IBNG-4: poor and IBGN-5: bad); B5: IBMR. **y'** is the reduced value of $\mathbf{y = n_{ab,ij}} / N_{i,}$ where $n_{ab,ij}$ is the number of appearances of the pressure category a in the status class b, for the given biological index i in the status class j and N_{ij} is the number of patterns obtained for a given biological index i (here IBGN) in the given status j; *the green squares on the bar charts for MOOX and PHOS recall that, for these pressure categories, the status class good was removed before patterns were extracted because of their dominance in each context.*

The largest number of patterns was obtained in the bad status context, which had the fewest sequences: 54% of patterns in the context "bad IBGN" versus less than 1% of sequences, and 70% of patterns in the context "bad IBMR" versus 1% of sequences. This was not the case for all the indices. For instance, we obtained 27% of patterns in the context "high IBD" versus 63% of sequences. That is why we chose to study the two indices IBGN and IBMR in more detail.

In patterns, pressure NITR appeared in only two statuses: good and moderate, even when the two biological indices were in poor or bad statuses. Its good status was gradually replaced by the moderate status when the status of the biological index worsened, but the proportion of good NITR remained high, especially in IBMR in bad status.

The pressure PEST appeared only in two statuses: poor and bad, even when the two biological indices were in high, good or moderate statuses. The biological status of IBMR worsened in both poor and bad statuses when the PEST status changed from poor to bad. This was not the case in IBGN where PEST in bad status was only found in good and moderate biological statuses.

The pressure MPOR in the moderate status appeared in IBGN, in poor and even more so in bad biological status. It appeared only in poor status in IBMR, which corresponds to a poor and a bad biological status.

The pressure PHOS appeared most frequently in high status which corresponds to high and good statuses in the two biological indices. The moderate status of this pressure was rarely observed in IBGN in poor status or in IBMR in bad status. It will be recalled that PHOS in good status was removed before extraction of the patterns of the chosen configuration because of their dominance in each context.

The pressure MOOX appeared most frequently in moderate status when the two biological indices were in moderate, poor and more particularly in bad statuses. The poor status of this pressure category was rarely found in IBGN in bad status. Their high status was rarely observed in IBGN in high status and in IBMR in high and good statuses. It will be recalled that MOOX in good status was removed before extraction of the patterns of the chosen configuration because of their dominance in each context.

Patterns were selected and analysed per context according to the following criteria: 1) highest frequency (f), 2) highest emergence (E), 3) highest complexity (C), 4) highest scarcity and 5) item diversity. Table **8** lists the first five patterns for each context except for bad IBMR, for which we included six patterns in order to have the pressure category AZOT at least once.

Table 8: Selection of five first patterns generated for all the contexts of the French biological indices IBGN and IBMR according to their frequency (f), emergence (E), complexity (C) and scarcity (S) and the result (f x C x S +E); The dominant context is the context in which the pattern is the most frequent; the last row indicates the items in each pattern; configuration: [area: France, period: 2007-2013, table of thresholds: SEQ, time-length: 24 months, minimum frequency: 0.6]; in bold: first pattern with no micro-pollutants.

Pattern	Generating Context	Dominant Context	f	E	C	S	f X C X S + E	Items
229	high IBGN	high IBD	**0.609**		**0.150**	**0.800**	**0.073**	**3 high PHOS**
228	high IBGN	high IBMR	0.604		0.150	0.640	0.058	3 poor PEST
221	high IBGN	high IPR	0.660		0.050	0.600	0.020	1 high MOOX
223	high IBGN	bad IBMR	0.619		0.050	0.160	0.005	1 bad PEST
220	high IBGN	bad IBMR	0.651		0.050	0.080	0.003	1 good NITR
307	good IBGN	bad IBMR	**0.637**		**0.100**	**0.360**	**0.023**	**2 good NITR**
303	good IBGN	high IBD	0.636		0.050	0.560	0.018	1 high PHOS
308	good IBGN	bad IBMR	0.624		0.100	0.240	0.015	2 bad PEST
305	good IBGN	bad IBMR	0.719		0.050	0.080	0.003	1 good NITR
304	good IBGN	bad IPR	0.738		0.050	0.000	0.000	1 good ACID
240	moderate IBGN	bad IBMR	0.629		0.200	0.640	0.080	4 poor PEST
239	moderate IBGN	bad IBMR	0.671		0.150	0.480	0.048	3 bad PEST
236	moderate IBGN	poor IBMR	**0.633**		**0.100**	**0.560**	**0.035**	**2 moderate NITR**
237	moderate IBGN	bad IBMR	0.605		0.100	0.360	0.023	2 good NITR
232	moderate IBGN	bad IBGN	0.675		0.050	0.440	0.015	1 moderate MOOX
254	poor IBGN	high IBMR	0.617		0.150	0.640	0.059	3 poor PEST
251	poor IBGN	poor IBMR	0.613		0.100	0.960	0.059	2 moderate MPOR
253	poor IBGN	bad I2M2	**0.613**		**0.100**	**0.560**	**0.034**	**2 moderate NITR**
242	poor IBGN	bad I2M2	0.613		0.050	0.920	0.028	1 moderate AZOT
244	poor IBGN	high IBMR	0.637		0.050	0.760	0.024	1 moderate PHOS
300	bad IBGN	bad IBGN	0.600	2.100	0.400	0.960	2.330	5 moderate MPOR, 1 moderate NITR, 1 moderate MOOX
266	bad IBGN	bad IBGN	0.600	1.719	0.250	0.960	1.863	4 poor MPMI & 1 moderate NITR
270	bad IBGN	bad IBGN	0.600	1.260	0.200	0.960	1.375	3 poor PEST & 1 moderate MOOX
289	bad IBGN	bad IBGN	0.600	1.273	0.100	0.960	1.331	**1 poor MOOX & 1 moderate NITR**
259	bad IBGN	bad IBGN	0.667	1.215	0.150	0.960	1.311	3 bad PEST
324	high IBMR	high IBMR	0.637	1.058	0.200	0.880	1.170	4 poor PEST
325	high IBMR	high IBD	**0.612**		**0.200**	**0.880**	**0.108**	**4 high PHOS**
323	high IBMR	bad IBMR	0.602		0.150	0.760	0.069	3 good NITR
316	high IBMR	high IPR	0.619		0.100	0.880	0.054	2 high MOOX
320	high IBMR	bad IBMR	0.604		0.100	0.240	0.015	4 bad PEST
745	good IBMR	high IBMR	0.625		0.150	0.640	0.060	3 poor PEST
746	good IBMR	bad IBMR	0.621		0.150	0.480	0.045	3 bad PEST
738	good IBMR	high IPR	**0.667**		**0.050**	**0.600**	**0.020**	**1high MOOX**
739	good IBMR	high IBD	0.643		0.050	0.560	0.018	1 high PHOS
735	good IBMR	bad IBMR	0.613		0.050	0.440	0.014	1 moderate MOOX
340	moderate IBMR	bad IBMR	0.6478		0.200	0.640	0.083	4 bad PEST
338	moderate IBMR	poor IBMR	**0.608**		**0.150**	**0.800**	**0.073**	**3 moderate NITR**
334	moderate IBMR	bad IBMR	0.637		0.100	0.480	0.031	3 poor PEST
335	moderate IBMR	high IBMR	0.659		0.100	0.360	0.024	2 good NITR
328	moderate IBMR	bad IBMR	0.704		0.050	0.440	0.016	1 moderate MOOX
410	poor IBMR	poor IBMR	0.604	1.228	0.300	0.960	1.402	4 bad PEST, 2 moderate NITR
363	poor IBMR	poor IBMR	**0.6923**	**1.119**	**0.150**	**0.960**	**1.219**	**2 moderate NITR, 1 moderate MOOX**
360	poor IBMR	poor IBMR	0.625		0.150	0.760	0.071	3 good NITR
421	poor IBMR	poor IBMR	0.604		0.100	0.960	0.058	1 moderate NITR, 1 moderate PAES
419	poor IBMR	poor IBMR	0.604		0.100	0.920	0.056	1 moderate NITR, 1 good ACID
623	bad IBMR	bad IBMR	0.619	4.647	0.500	0.960	4.944	2 bad PEST, 4 good NITR, 2 moderate MOOX, 2
689	bad IBMR	bad IBMR	0.619	3.714	0.950	0.960	4.279	9 bad PEST 4 good NITR 3 moderate MOOX 3 poorMPOR
564	bad IBMR	bad IBMR	**0.619**	**3.714**	**0.350**	**0.960**	**3.922**	**3 good NITR, 1 moderate MOOX, 3 moderate PAES**
674	bad IBMR	bad IBMR	0.619	3.200	0.600	0.960	3.557	3bad PEST 4 good NITR 2 moderate NITR, 3 poor MPOR
494	bad IBMR	bad IBMR	0.619	2.758	0.500	0.960	3.055	5 good NITR, 4 moderate MOOX, 1 good MINE
442	bad IBMR	bad	0.61905	2.207	0.250	0.960	2.356	1 moderate AZOT, 1 good MINE, 3 good NITR

Among the 50 selected patterns, four contexts had patterns with no null emergence: IBMR in high, poor and bad statuses and IBGN in bad status. These were the only ones to be dominant contexts with 4 other contexts with other indices than IBGN and IBMR (high IBD, high IPR, bad I2M2 and bad IPR). The other generating contexts, with IBGN and IBMR indices, were never dominant for any given pattern. Only the patterns of four generating contexts had a complexity higher than 0.25 (in this configuration, the biggest pattern had 20 items and C=0.25 corresponded to a pattern with five items): the patterns for IBMR in high, poor and bad statuses and IBGN in bad status. Patterns for only four contexts had a scarcity less than 0.76 (in this configuration, S=0.76 was obtained for a pattern found in more than six contexts): the patterns for IBMR and for IBGN in good and moderate statuses.

In each context, we performed detailed analysis of the first pattern according to the selected classification, and the first one with macro-pollutants. The first pattern for high IBGN (pattern 229, Figure 25) and the second for high IBMR (pattern 325, Figure 26) showed that the pressure PHOS remained stable in high status, during respectively three and four successive measurements. Pattern 738 (Figure 26), obtained for the generating context high IBMR, was the only one with another pressure category in high status, i.e. one item with high MOOX. The stability of pressure PEST in poor or bad statuses, measured successively three times (patterns 228, 334, 745 in Table 8, pattern 254 in Table 8 and Figure 25, patterns 259 and 746 in Table 8) or four times (patterns 324 in Table 8 and Figure 26, pattern 320 in Table 8 and patterns 240 in Figure 25 and 340 in Figure 26), was not discriminating with respect to a specific context. The pressure category "good NITR" did not appear to be discriminating; as whatever the context, it appeared: once in high and good IBGN (patterns 220 and 305, Table 8), twice in good, moderate IBGN (patterns 307, Figure 25 and pattern 237, Table 8) and moderate IBMR (pattern 334, Table 8), three times in moderate IBGN and poor IBMR (patterns 323 and 360,Table 8). Moderate NITR appeared more specifically in poor and bad contexts in the two indices (patterns 236, 253, Table 8, and 338, Table 8 and Figure 26). Pattern 289 (Figure 25) was specific to bad IBGN (E= 1.27333 and S=0.96): meaning that only one poor MOOX was measured before or after one moderate NITR. Pattern 300 (Figure 25) was also specific to bad IBGN (E= 2.1 and S=0.96): meaning that pollution with moderate

MPOR measured five times remained stable, and two moderate NITR and one moderate MOOX were measured after or before one these measurements. In pattern 442 (Figure 26), only one measurement of moderate AZOT between one good MINE and three good NITR were observed before a bad IBMR. Pressure MPMI appeared only in bad IBGN contexts, e.g. in pattern 266 (Figure 25) composed of four poor MPMI and one moderate NITR. In the bad IBMR context, patterns 623 and 564 (Figure 26) represented multi-stress conditions: moderate MOOX, and moderate PAES for the first, with added bad PEST for the second.

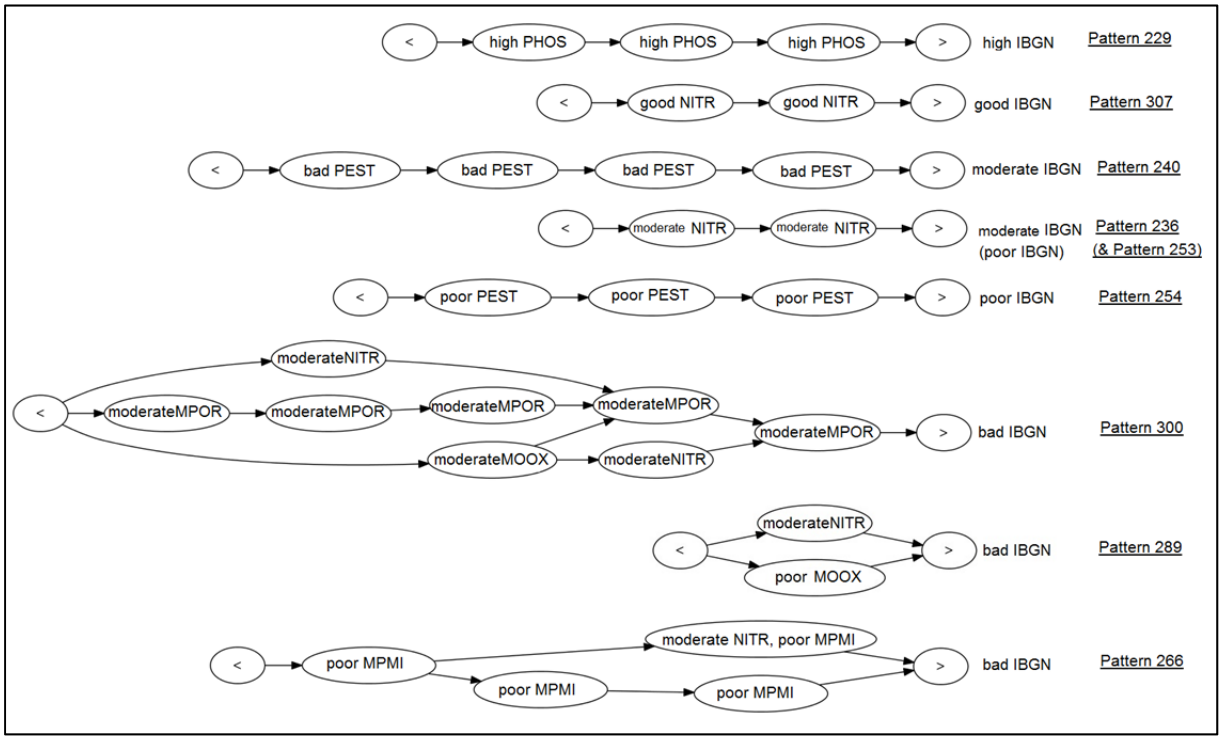


Figure 25: Nine major patterns extracted from the five contexts of IBGN in the configuration: [area: France, period: 2007-2013, table of thresholds: SEQ, time-length: 24 months, minimum frequency: 0.6]

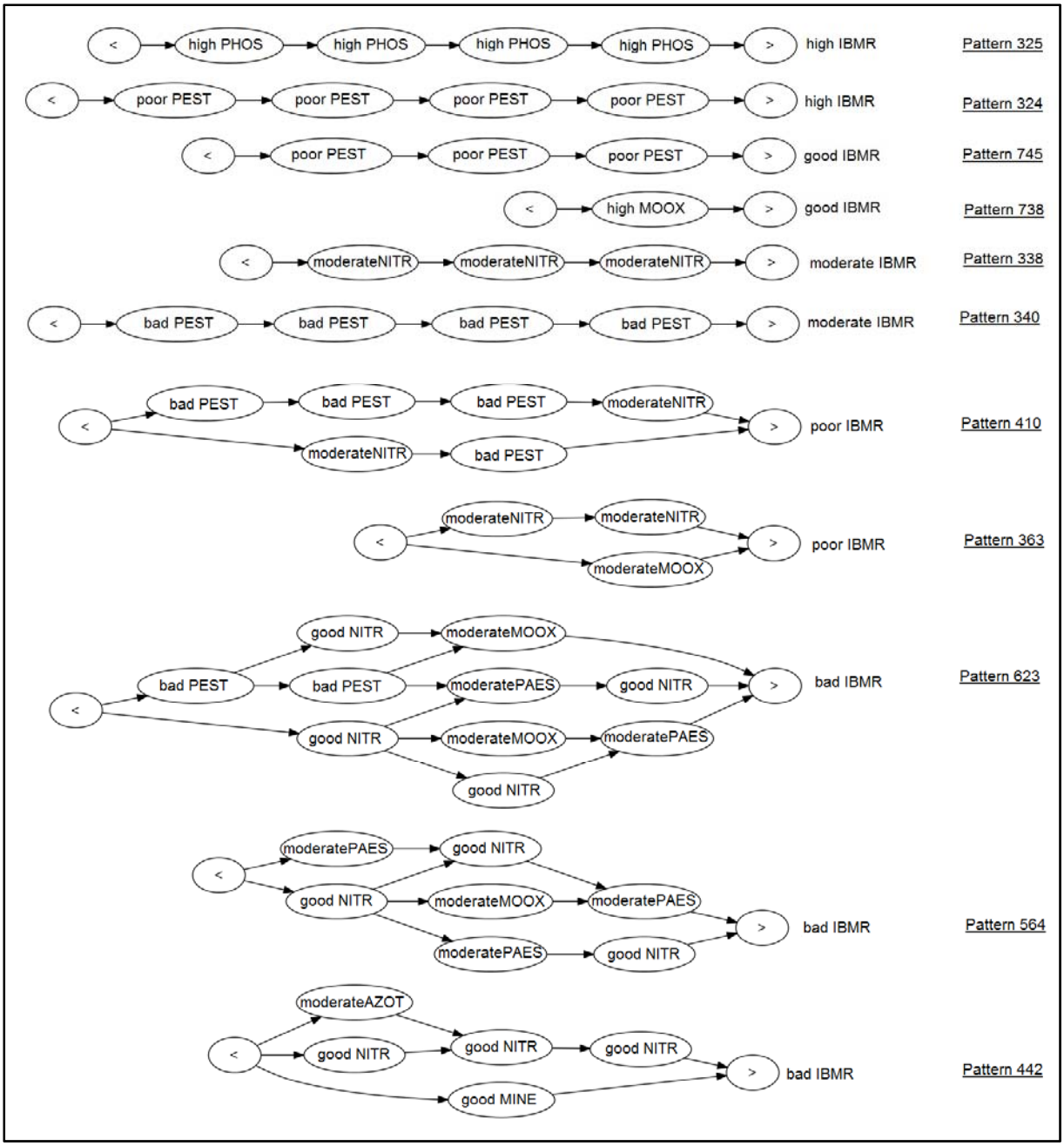


Figure 26: Eleven major patterns extracted from the five contexts of IBMR in the configuration: [area: France, period: 2007-2013, table of thresholds: SEQ, time-length: 24 months, minimum frequency: 0.6]

2.6 Discussion

A promising tool to reveal temporal sequences preceding a biological status

We analysed a large dataset in an innovative way using an efficient, specific but flexible tool which extracts and selects the most relevant temporal patterns of physico-chemical variables that precede a target representing a biological index. Fabrègue et al. (2014) were the first to show the promising potential of patterns to describe sequential data concerning quality in river ecosystems. We continued using a bigger dataset (here 1,146,544 physico-chemical data and 24,593 biological data) and by multiplying the possible configurations. The PRESTOR tool was specifically applied to extract patterns based on different criteria (time-length, minimum frequency, etc.) directly from hydro-ecosystem databases. We observed that the number of patterns increased with an increase in time-length and with a decrease in the minimum frequency or in the number of sampling sites, but these are trends without real correlation. Whereas extraction is stable and rapid, many tests have to be conducted to tune the parameters, and, like in multivariate analysis, we first had to remove the dominant pressure categories in order to analyse all the data in detail using a top-down approach. As shown by Serrano Balderas et al. (2017), selecting features for data reduction is an essential data preprocessing technique to ensure the final results are valid.

To proceed from a potentially useful to an operational method using PRESTOR, we propose a double process before and after extraction to control the number of patterns extracted and to identify the most relevant ones. Analysis of the 809 patterns we extracted is irrelevant for an expert both timewise and for the final results. All patterns are not equally useful for interpretation. Some are nested in others, for example, a pattern of three successive “poor PESTs” included in another one of four successive “poor PESTs” (Figure 26). After extraction, the three metrics (complexity (C), scarcity (S) and emergence (E)) help select the most relevant patterns among hundreds or even thousands of patterns.

The process generated more patterns for the smaller sub-datasets. This bias is shared by other methods as well as by data mining and statistical methods, which is why the pre and post-processing are recommended particularly for environmental data mining (Gibert et al., 2008). Here, there can be many patterns in a given context for different reasons: 1) few sequences were available: in our case, IBMR represented 8% of sequences and extracted 53% of patterns, because the threshold of minimum frequency is easily reached; 2) data were heterogeneous: input data for IBMR were more heterogeneous than those for IBD, which represented the maximum number of sequences (27%) and extracted only 18% of patterns; 3) data were numerous but homogeneous, as were data for high IBD in the input data. This problem was amplified by unbalanced data distribution. The advantage of this dataset is that it is the complete set of French data from a survey network covering six years, but we nevertheless had to deal with the real distribution of these data: more major parameters than minor pollutants, a different number of sequences available for each index and fewer sequences for bad ecological status.

The selection of the same number of the most relevant patterns for a given biological index with a given status, thanks to the three metrics, limits this bias. Before focusing on the most relevant patterns, to be sure the method is efficient, we studied the complete batch of patterns extracted for a time-length of two years.

Feedback on the pressure categories analysed and the quality of data

In our analysis, six types of pressure categories can be identified according to the analysis of the raw data in the first case and to the patterns extracted in the other cases.

The first type were several pressure categories considered to remain stable in a single status throughout the French territory, i.e. temperature (TEMP), acidification parameters (ACID) and effect of eutrophication (EPRV) in high status and polycyclic aromatic hydrocarbons (HAP) in moderate status. For temperature, in the SEQ (op. cit) or in the WFD guide (op.cit.), the quality class is defined by higher temperatures

and the temperature difference between high and good class thresholds is 1.5 °C. This difference may correspond to the 1.4 °C to 1.7 °C increase observed by Durance and Ormerod (2007) in streams in Wales between 1981 and 2005 or to the forecasts of 1- 9 °C announced by Webb (1996) for 2050. But the mechanisms which control freshwater temperature behavior in the context of global climate change are complex, in particular because temperature increases have been observed in winter while summer temperatures have remained relatively stable (Webb et al., 2008). This could explain why, in the French dataset, temperatures remained stable in high status, and underlines the need to change the way we define this pressure category. In Europe and North America, acidification of freshwater decreased throughout the 1980s and 1990s thanks to the reduction of atmospheric emissions of sulphur dioxide (Stoddard et al., 1999). Therefore, pressure due to acidification is now localized rather than widespread, which is probably why it is difficult to observe this pressure in a study conducted at national scale. The quality class for the effect of eutrophication (EPRV), in the SEQ (op. cit), is generally defined for water combining basic pH and oversaturated in oxygen. In our dataset, this pressure category was mainly in high quality based on this definition, which is why it was not included in the analysis of patterns. Again, HAP are ubiquitous contaminants to a moderate degree in our study area, as also shown by Motelay-Massei et al. (2004) in the soils of the Seine River Basin.

The five following pressure categories are proposed according to the patterns extracted from the complete dataset with a time length of two years.

The second pressure category was measured in different statuses but considered as dominant in only one status. There were macro-pollutants in high or good quality statuses: mineralisation (MINE) in high status, suspended matter (MES), oxidisable organic matter (MOOX), nitrogen excluding nitrate (AZOT) and phosphorus matter (PHOS) in good status. We had to remove them from the final extraction in order to see other pressure categories, but they can be considered as regularly measured in each site.

The third type includes absent pressure categories or categories with a weak presence in the patterns. Thus polychloro-biphenyls (PCB) did not appear in any patterns. This pressure category was the one with the lowest number of measurements in the dataset (completeness rate: 33% versus at least or more than 50% for the other micro-pollutants). Teil et al. (2004) estimated that the pressure exerted by PCB in the Seine River Basin to be beyond the global national estimation of industrial inputs, but we did not observe it at the national scale. Pressure categories ACID and MINE in other classes than the high one accounted for less than 4% of all appearances of pressures.

The fourth type of pressure category was dominant in patterns but in no or in only poorly discriminating pressure categories regardless of biological status: pesticides (PEST) was only found in poor or bad statuses in the patterns we extracted. This pressure category was the one in which the number of measurements of micro-pollutants was the highest in the dataset (completeness rate: 61%). According to the Millenium Ecosystem Assessment Programme (2005), pesticides are one of the major stressors in freshwater ecosystems. The patterns extracted revealed the spatial and temporal dominance of this pressure category in the French dataset. After nitrates, it was the second most dominant category in all the patterns extracted, but was found in all biological statuses from high to bad. There was no graduation either in status or in the number of appearances in the patterns extracted for the IBGN index. Conversely, there was a small gradation in status and in the number of appearance in patterns extracted for the IBMR index: poor PEST appeared only in patterns extracted for high to moderate IBMR, whereas patterns extracted for poor and bad IBMR contained only bad PEST. The proportion of herbicides (45%) versus insecticides (31%) in the parameters included in PEST in SEQ (op.cit.) could partly explain the small graduation observed in patterns extracted for IBMR. Even if macroinvertebrates are known to be the most sensitive groups to pesticides in freshwater communities (Schäfer et al., 2012), we observed different responses of the two indices we tested, IBGN sensitivity was higher than that of I2M2 (Mondy et al., 2012).

The fifth category grouped distinguishing pressure categories for a given biological status in all biological indices: nitrates (NITR) in good and moderate status, phosphorus (PHOS) in high and moderate status, oxidizable organic matter (MOOX) in high, moderate and poor status, and other organic hydrocarbons (MPOR) in moderate and poor status. Like in recent studies dealing with the effects of multiple stressors on freshwater biodiversity, we found nutrients among the parameters that have the most impact (Lemm and Feld, 2017; Stendera et al., 2012) along with organic matter (Comte et al., 2010; Villeneuve et al., 2015). We are the first to underline the importance of other organic hydrocarbons (MPOR) in patterns especially for IBGN, IPR and IBMR contexts.

The sixth pressure category was discriminating for a given index in a given biological status: heavy metals (MPMI) in poor status. This pressure category is specific to patterns extracted for IBGN index in bad status. It was missing in patterns for I2M2 the other French index based on macroinvertebrates whereas Mondy et al. (2012) found a better correlation between this stressor with I2M2 than with IBGN.

A specific biological response to a specific pressure category?

Except for heavy metals (MPMI), none of the pressure categories was specific to one biological index in the patterns extracted from the French dataset we used. Many authors who used large national indices and several biological indices reported that all biological groups responded firstly to macro-pollutants in general (Dahm et al., 2013; Marzin et al., 2012), particularly to organic matter and nutrients (Villeneuve et al., 2015). Haury et al. (2006) showed that the IBMR was sensitive to trophic disruption but also to heavy organic pollution. Nutrients did not appear to be specific to patterns extracted for IBMR or IBD indices. However high and good IBMRs (generating contexts) were associated with High PHOS and MOOX or good NITR. PHOS was also the dominant pressure category for patterns extracted for IBD index (dominant context). Nevertheless, excess nutrients are known to have indirect effects on aquatic organisms, especially on macroinvertebrates (Dolédec et al., 2006; Lemm and Feld, 2017). The pressure acidification (ACID) was well represented (18%) in the patterns extracted for the IBD index, whereas Larras et al. (2017) found

a weak correlation between this pressure and diatoms, but their approach using life history traits differs substantially from our approach. Fish are known to respond less well to physico-chemical pressures (Dahm et al., 2013; Marzin et al., 2012; Villeneuve et al., 2015). Indeed, the patterns extracted for IPR contained the lowest number of pressures (6/10). Adding hydrological and hydromorphological pressures in the dataset to produce new patterns appears to be a promising solution. Few authors who used large datasets tested biological responses to all micro-pollutant categories, except Larras et al. (2017) and Mondy and Usseglio-Polatera (2013) respectively for diatoms and macroinvertebrates using the same French dataset. These autors found good biological responses for pesticides but poor responses for PAH, heavy metals (MPMI) and other organic hydrocarbons (MPOR), whereas the last pressure category was widely represented in the patterns we extracted in our study.

Gap between physico-chemical statuses and biological statuses

In the differentiating pressure categories, physico-chemical statuses were often better distinguished than the biological statuses in the patterns. Oberdorff and Hughes (1992) already observed a gap between the precursor of the IPR index and SEQ results in the Seine catchment. Except for suspended matter (PAES) and pesticides (PEST) found in bad status in some patterns, the worst status found for other pressure categories was poor status. Defining class boundaries is a critical step in the design of methods of assessment (Birk et al., 2012). Using new French thresholds (MEEM, 2012) to discretize physico-chemical data, we obtained less significant results. This can be partly explained by the fact that these pressure categories include more parameters, but also by the change in the boundary between the good and moderate threshold for nitrates in the nutrient pressure category, which increased from 10 mg/L in SEQ (MEDD and AE, 2003) to 50 mg/L in the WFD guide (MEEM, 2012). This is at least the second time that this threshold has been revised upwards: 50 years ago, Nisbet and Verneaux (1970) set the boundary for excess nitrates in French freshwater at 3 mg/L. Although good-bad thresholds for micro-pollutants were discussed at the European level for the WFD guide, there was no harmonization of the thresholds for macro-pollutants, and the literature comparing

them in Europe is really poor. Current French thresholds are among the highest in Europe probably because the majority of countries take the mean or the median as the annual value while France takes the 90th percentile. The French good-moderate thresholds compared to the European ranges are 1) for NH_4^-: 0.5 mg/L in the range [0.05-1.6 mg/L] (Claussen et al., 2012), 2) for O_2: 6 mg/l in the range [6-10 mg/L] - but the higher values are used by countries specifying different thresholds for different stream typologies (based on size, climate, geology, geographical location), which is not the case in France - (Claussen et al., 2012), 3) for orthophosphates: 0.5 mg/L in the range [0.05-1 mg/L] (Arle et al., 2016), 4) for nitrogen adding the French good-moderate thresholds for Kjeldahl nitrogen, nitrites and nitrates, the boundary is 4.5 mg/L for SEQ (MEDD and AE, 2003) and 14 mg/L for WFD guide (MEEM, 2012) versus the European range [0.7-10 mg/L] (Claussen et al., 2012).

The stronger reactions of living organisms evidenced by worse states than their physico-chemical status, can also be explained by possible synergism between stressors and/or an additive impact of other stressors related to hydromorphology or hydrology. According to Lemm and Feld (2017), combined nutrient and hydromorphological stress can strengthen an individual's reactions to each single stressor for different biological traits.

For the same reasons, synergism between stressors and/or an additive impact of other physical stressors, extreme ecological status classes are the easiest to define (Birk et al., 2012). This is surely why in the configuration we chose, or in other configurations, intermediate biological statuses i.e. good and moderate, never displayed specific patterns. Conversely, extreme statuses, high and bad, often displayed specific patterns. For the poor status, in the configuration we chose, there were some specific patterns for IBMR, but not for IBGN.

Characteristic physico-chemical successions for high and bad biological statuses

Patterns of high biological status were characterised by consistently high physico-chemical status especially for phosphorus (PHOS) measured three or four times (patterns 229, Figure 25 patterns 324 and 325, Figure 26), and probably in the consistency of the pressure categories we had to remove before the extraction: mineralisation (MINE) in high status, suspended matter (MES), oxidizable organic matter (MOOX), nitrogen excluding nitrate (AZOT).

We found two types of patterns in bad biological status: (1) a chronic multi-pressure one, in which pressure categories such as nitrates, pesticides and other organic hydrocarbons, in moderate, poor or bad status, repeated themselves several times over time, or (2) a single occurrence of a degraded pressure category, such as one moderate nitrogen excluding nitrate, or one poor oxidizable organic matter, among other pressure categories in good status (respectively pattern 289, Figure 25 and pattern 442, Figure 26).

Studies of the effect of changes in pressure categories over time are scarce or limited to general trends in the case of degradation such as the hydraulic management of large rivers (e.g. Fruget et al., 2001; Trémolières, 1994) or more recently in the case of restoration (e.g. Meyer et al., 2013; Staentzel et al., 2017). However databases storing temporal monitoring data on rivers do exist. Most studies used data on each site on each sampling occasion as timeless information to increase the gradient of measurements before timeless treatments, because the methods used were not able to incorporate the temporal dimension (D'heygere et al., 2006; Dahm et al., 2013; Larras et al., 2017; Marzin et al., 2012; Mondy and Usseglio-Polatera, 2013; Villeneuve et al., 2018). This is why the temporal patterns extracted by PRESTOR are innovative and offer new opportunities to explore data on rivers.

2.7 Conclusion

Applying a new data mining method for river ecological assessment required iterative collaboration between computer scientists and experts in the domain to (1) adapt the method to a particular question and to the specific format of the data, (2) analyse the complete results and not only those that "matched well" to ensure the methods are efficient, (3) to propose selection criteria for the evaluation of patterns, (4) to select the most important results particularly for data mining methods which have led to exponential increase in results. PRESTOR was implemented specifically to extract patterns from a hydro-ecosystem database. The operator can choose different criteria such as the time-length of patterns or their minimum frequency. To check the efficiency of the method, we analysed the 809 patterns extracted from data collected all over France with a time-length of 24 months. We propose three metrics, complexity, scarcity and emergence, to select significant patterns according to their ecological status.

Water managers need simple and transparent methods. Selected patterns extracted by PRESTOR are easily readable and match managers' needs. They could be used at large scale and as such, considered as a holistic approach, the kind of approach that is urgently needed (Demars et al., 2012; Stendera et al., 2012). In a multi-pressure environment (Reyjol et al., 2014), new applications could be using the dataset including hydromorphological and physico-chemical pressures to extract patterns. In this study, we extracted patterns over a large territory (France), but the method could equally be applied to a small territory, for example, a hydro-eco-region, and could extract more discerning patterns. Patterns, highlighting a sequence of alteration events before an observed biological response, are a promising solution to disentangle the effects of different pressures, especially in a context of multi-stress conditions.

2.8 Acknowledgments

This work was funded by the French National Agency of Biodiversity (research support agreement n°460, December 23rd, 2014).

Appendix 1: Set of thresholds used by SEQ (MEDD and AE, 2003) for the physico-chemical parameters classified according to pressure categories *(here the threshold is always excluded from the better class, e.g.: for nitrates, the threshold between the high and good classes is 2 mg/L, therefore if [NO_3^-]=2 mg/L, pressure category NITR is good)*

Appendix 2: Set of thresholds used by the WFD guide (MEEM, 2012) for physico-chemical parameters classified according to pressure categories *(here for TEMP, ACID, BILO2, NUTRI: the threshold is always excluded from the better class, e.g.: for nitrates, the threshold between the high and good classes is 10 mg/L, therefore if [NO_3^-]=10 mg/L, pressure category is good; in contrast, for POSPE and SDP the threshold is always included in the better class, e.g.: for arsenic, the threshold between the high and bad classes is 4.2µg/L, therefore if [As]=4.2µg/L, pressure category is high)*

Avertissement: nous avons choisi de ne pas mettre dans cette thèse les annexes 1 et 2 de l'article présenté ici. En effet, l'annexe 1 correspond aux seuils du SEQ-eau (MEDD et AE, 2003) et l'annexe 2 aux grilles DCE de 2012 (MEEM, 2012). Ce sont deux documents facilement consultables en France.

CHAPITRE IV: Existe-t-il des différences entre motifs extraits par indice biologique en fonction des longueurs de séquences considérées?

Nous avons développé le programme, PRESTOR, qui permet à un opérateur d'extraire des motifs pour un indice biologique dans un état donné, le contexte. L'opérateur peut choisir différents critères d'extraction, dont le territoire (la France entière ou une Hydro-Eco-Région (HER) donnée) et la longueur des séquences (stations-dates), c'est-à-dire l'intervalle de temps précédant l'état biologique. Dans le chapitre précédant nous avons montré quelles étaient les altérations les plus fréquentes dans les motifs extraits pour l'IBGN et l'IBMR pour une longueur de séquences de 24 mois.

Les cinq indices biologiques dont nous disposons sont basés sur quatre groupes d'êtres vivants – invertébrés, poissons, diatomées, macrophytes – qui n'ont pas les mêmes cycles de vie, les mêmes longévités potentielles, ni les mêmes réponses aux différentes pressions (Lafont, 2001; Marzin, 2013). Les organismes devraient être sensibles pour une période précédant leur durée de développement qui devrait être cohérente avec leur durée de vie. Aussi la question posée dans ce chapitre est-elle la suivante : **les motifs extraits pour différentes longueurs de séquences sont-ils différents en fonction des indices biologiques?** Sur le même jeu de données national, nous analyserons les motifs extraits pour les différentes durées 3, 6 12, 18 et 24 mois, pour les cinq indices biologiques: l'IBGN, l'I2M2, l'IBMR, l'IBD et l'IPR. Nous verrons également si nous obtenons des données comparables en utilisant les altérations et les seuils des grilles du SEQ-eau et celles dites DCE.

1 Matériel et méthodes

Le Tableau 6 synthétise les caractéristiques des extractions réalisées sur les cinq longueurs de séquences choisies. Les extractions ont été faites pour les deux grilles de discrétisation: le SEQ-eau et celles de la DCE (MEEM, 2012). Comme dans le chapitre III (2.5.2), nous avons enlevé avant les extractions les altérations présentes à 80% dans les données initiales – c'est le cas pour le SEQ-eau des altérations température (TEMP) et acidité (ACID) et effets proliférations végétales (EPRV) de très bonne qualité et les hydrocarbures aromatiques polycycliques (HAP) de qualité moyenne ; pour les grilles DCE, c'est le cas des altérations température (TEMP), acidité (ACID) en très bon état. Nous avons également enlevé les altérations dont le cumul représentait plus de 90% des items à la première extraction – c'est le cas pour le SEQ-eau des altérations minéralisation (MINE) de très bonne qualité, des particules en suspension (PAES), des matières organiques et oxydables (MOOX), des matières azotées (AZOT) et des matières phosphorées de bonne qualité. Pour répondre à la question posée dans ce chapitre, nous avons également choisi d'enlever l'altération nitrates (NITR) de bonne qualité pour les grilles SEQ-eau ; et l'altération nutriments (NUTRI) en bon état pour les grilles DCE. En effet dans l'ensemble des extractions réalisées, ces altérations représentent plus de 20 % des items, quelle que soit la longueur de séquences choisie, exceptée pour 6 mois, avec la grille SEQ-eau (15%).

Tableau 6 : Caractéristiques des extractions réalisées

N° extraction	Paramètres d'extraction						Nb stations	Nb motifs
	Période	Territoire	Grilles de seuils	Intervalle de temps	Fréquence minimale	Altérations non prises en compte		
1	2007-2013	France	SEQ	3	0.30	TEMP=Bleu /ACID=Bleu/ EPRV=Bleu /PAES=Vert/ MOOX=Vert/ AZOT=Vert/ MINE=Bleu/ PHOS=Vert/ HAP=Jaune/ NITR=Vert	1636	272
2				6	0.40			279
3				12	0.53			237
4				18	0.60			201
5				24	0.65			282
6			DCE	3	0.48	TEMP=Bleu /ACID=Bleu/ NUTRI=Vert		229
7				6	0.60			217
8				12	0.76			264
9				18	0.80			272
10				24	0.88			236

Nous ne pouvons pas avoir une fréquence minimale identique entre les différentes extractions : en effet si cette limite est trop faible pour les longues successions temporelles, nous atteignons les limites de l'algorithme et l'extraction ne peut aboutir. Ainsi avec le SEQ-eau à 18 mois, la fréquence minimale ne peut pas être inférieure à 0,6. A l'inverse, si cette limite est trop élevée pour de faibles durées, le nombre de motifs obtenus est trop faible. Aussi avons-nous choisi des fréquences nous permettant à la fois de rester dans les possibilités de l'algorithme et d'obtenir un nombre de motifs par extraction entre 200 et 300.

Nous avons utilisé la combinaison (P) des quatre mesures d'intérêt proposées dans le chapitre III (2.4.4) pour sélectionner cinq motifs caractéristiques par contexte par ordre décroissant de P (Equation 4).

Equation 4 : $$P = F \times C \times S + E$$

Où F: fréquence du motif (intervalle de valeurs [0-1])

C : complexité (« *complexity* ») du motif (intervalle de valeurs [0 ;1])

S : singularité (« *scarcity* ») du motif (intervalle de valeurs [0 ;1])

E : émergence (« *emergence* ») du motif (intervalle de valeurs [0 ;∞])

2 Résultats obtenus pour les grilles SEQ-eau

2.1 Caractérisation des ensembles de motifs obtenus par longueur de séquences

Comme décrit dans le chapitre III (2.6), le nombre de motifs obtenu par extraction est plus important pour les classes extrêmes, en particulier pour l'IBD en très bon état, et les l'IBGN et l'IBMR en mauvais état, quelle que soit la longueur des

séquences (Figure 27). Les motifs dominants sont trouvés dans ces seuls contextes, ainsi que dans le contexte IBMR médiocre. Les motifs totaux sont les plus nombreux pour le contexte IBGN en mauvais état pour les durées 3, 6, 12 mois (respectivement 35, 49 et 41), et pour l'IBMR en mauvais état pour 24 mois (74). Pour la durée 18 mois, les nombres de motifs les plus élevés sont équivalents pour les contextes IBGN en mauvais état (22), IBMR en mauvais état (21) et IBD en très bon état (20).

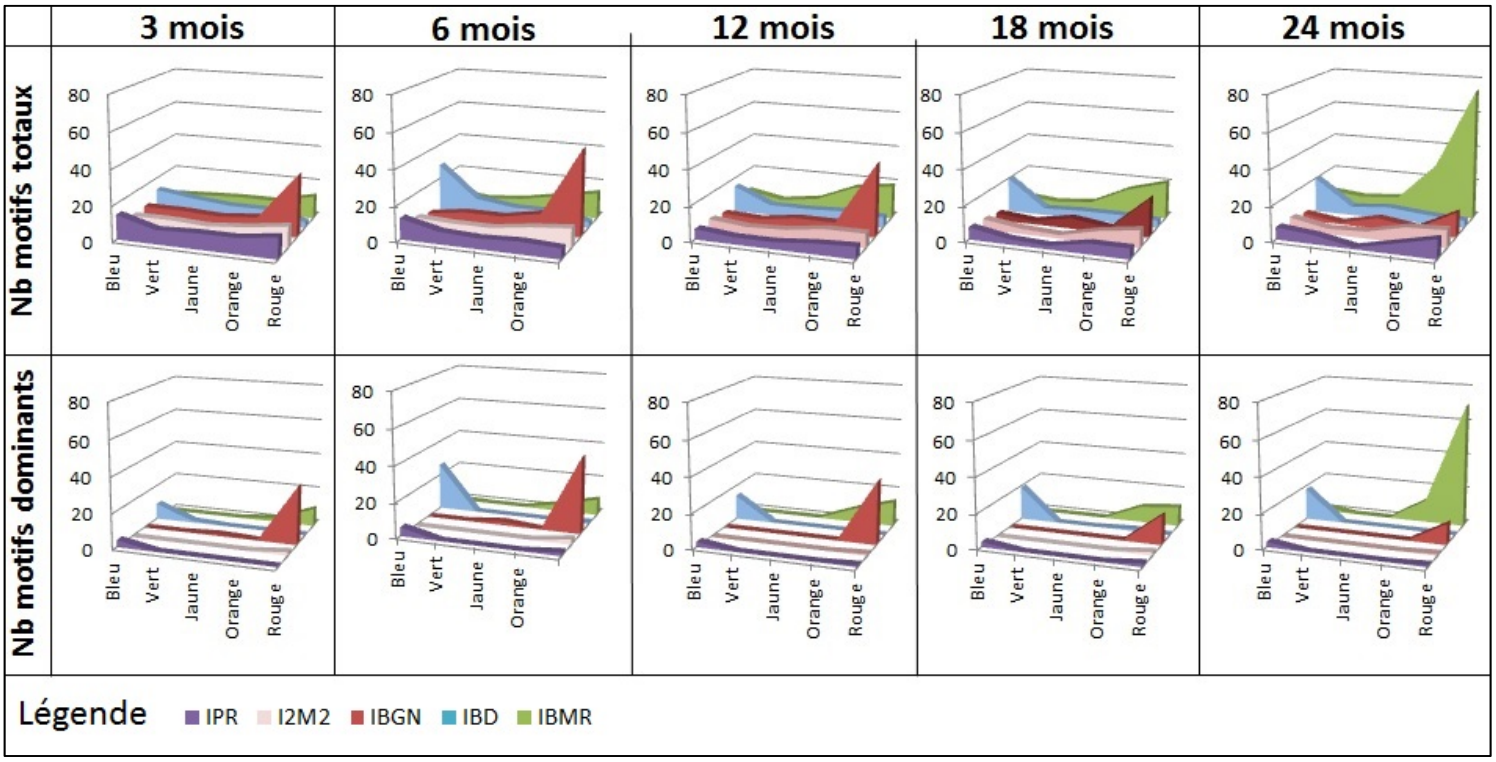


Figure 27 : Nombre de motifs et de motifs dominants par indice biologique et par classe de qualité du SEQ-eau (de 1, très bonne à 5 mauvaise) obtenu pour les extractions réalisées pour les longueurs de séquences 3, 6, 12, 18 et 24 mois

Le Tableau 7 fournit les principales valeurs de dispersion – médiane, minimum, maximum, 1er et 3ème quartiles et variance – des mesures d'intérêt F, C, S, E et de leur combinaison P calculées sur l'ensemble des motifs obtenus pour les extractions réalisées pour les cinq longueurs de séquences de 3, 6, 12, 18 et 24 mois. Sur la Figure 28 sont représentés les quartiles, les minima et maxima en fonction des cinq longueurs de séquences, par mesures d'intérêt et combinaisons.

Rappelons que les limites de l'algorithme nécessitent d'augmenter progressivement la fréquence minimale, ce qu'illustre la dispersion de la fréquence (Figure 28). Pour chaque extraction, les trois quart des fréquences des motifs extraits restent proches de cette fréquence minimale. Les variances de F restent faibles quelle que soit la longueur de séquences considérée [0,005 ; 0,006].

Tableau 7 : Variation – médiane, variance, écart-type, 1er et 3ème quartile, minimum, maximum – des valeurs des mesures d'intérêt F, C, S, E et de leur combinaison P calculées sur l'ensemble des motifs obtenus pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec le SEQ-eau

Mesures d'intérêt	Longueur séquences (mois)	MEDIANE	VARIANCE	ECART-TYPE	1ère QUARTILE	3ème QUARTILE	MIN	MAX
F	3	0,362	0,005	0,074	0,327	0,417	0,300	0,794
	6	0,461	0,006	0,077	0,421	0,517	0,400	0,845
	12	0,581	0,004	0,066	0,551	0,630	0,530	0,886
	18	0,656	0,004	0,062	0,619	0,700	0,600	0,900
	24	0,698	0,003	0,055	0,667	0,730	0,651	1,000
C	3	0,167	0,021	0,144	0,167	0,333	0,167	1,000
	6	0,520	0,022	0,147	0,320	0,960	0,000	0,960
	12	0,250	0,036	0,190	0,125	0,375	0,125	1,000
	18	0,167	0,042	0,205	0,167	0,375	0,167	1,000
	24	0,167	0,085	0,293	0,083	0,250	0,000	1,000
S	3	0,480	0,106	0,326	0,160	0,840	0,040	0,960
	6	0,083	0,111	0,334	0,083	0,167	0,083	1,000
	12	0,720	0,105	0,325	0,360	0,960	0,000	0,960
	18	0,760	0,104	0,324	0,520	0,930	0,000	0,960
	24	0,880	0,030	0,175	0,480	0,960	0,000	0,960
E	3	1,663	0,320	0,570	1,361	2,352	0,160	3,071
	6	1,529	0,364	0,607	1,245	2,072	0,083	3,538
	12	1,428	0,280	0,533	1,189	2,027	0,360	3,200
	18	1,257	0,063	0,252	1,164	1,379	0,520	2,117
	24	1,272	0,086	0,294	1,178	1,513	0,480	2,365
P=FCS+E	3	0,034	0,809	0,901	0,011	0,589	0,840	3,337
	6	0,037	0,938	0,971	0,018	1,509	0,000	3,965
	12	0,087	0,909	0,955	0,030	1,433	0,960	3,648
	18	0,149	0,596	0,774	0,057	1,400	0,930	2,501
	24	0,157	0,657	0,812	0,041	1,478	0,960	2,687

On pourrait s'attendre à une augmentation progressive de la taille des motifs avec la longueur des séquences. Cela est vérifié à 3 mois et 24 mois où les nombres d'items maximaux sont respectivement de 6 et 12. Mais ce n'est pas vrai pour les autres longueurs de séquences : le nombre maximal d'items est respectivement de

12, 8 et 6 pour 6, 12 et 18 mois. La durée de 6 mois apparait comme le meilleur compromis avec la fréquence minimale de 0,40 pour obtenir les motifs les plus complexes. Mais à l'inverse, ils sont peu singuliers, c'est-à-dire rarement spécifiques d'un seul contexte : trois quarts des valeurs de S sont inférieures à 0,167 alors qu'elles sont supérieures à 0,840 (3° quartile) pour les autres longueurs de séquences.

Globalement, l'émergence (E) des motifs diminue avec l'augmentation de la longueur de séquences. En revanche, bien que la médiane de la combinaison P des mesures d'intérêt augmente régulièrement avec la longueur des séquences (de 0,034 pour 3 mois à 0,157 pour 24 mois), les dispersions de ses valeurs sont équivalentes pour les longueurs de séquences exceptées pour 3 mois où le 3ème quartile est plus faible (0,589 contre 1,4 à 18 mois).

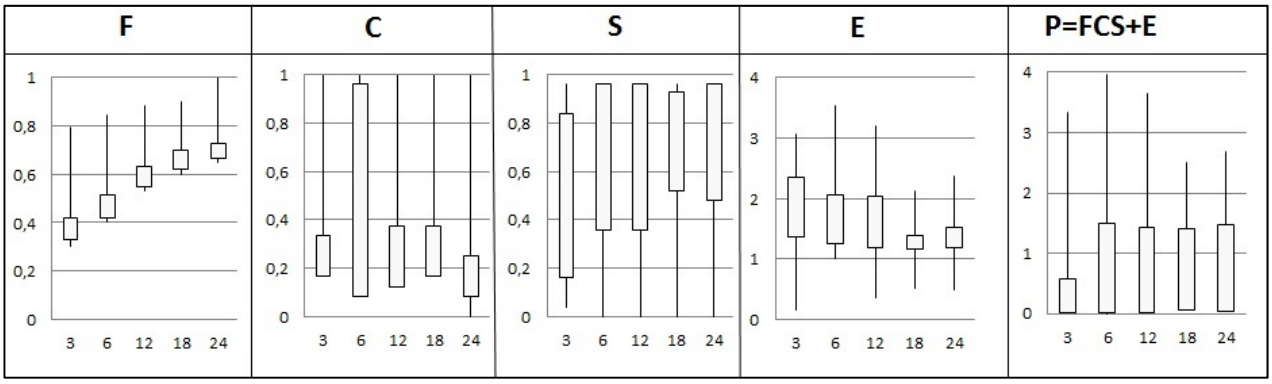


Figure 28 : Variation des mesures d'intérêt F, C, S, E et de leur combinaison P calculées sur l'ensemble des motifs obtenus pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec le SEQ-eau (les extrémités du rectangle représentent les 1ers et 3èmes quartiles, les extrémités des lignes les minima et maxima)

Les représentations graphiques des différentes mesures d'intérêt par indices biologiques et par classes d'état biologique, pour les différentes longueurs de séquences sont données en annexes 1 et 2 et dans la Figure 29 pour leur combinaison P. Les classes d'état biologique intermédiaire : bonne (2), moyenne (3), ont des motifs peu caractéristiques; les médianes de P sont très proches de zéro. Il s'agit en général de motifs non dominants (E=0), courts (faible C : médiane = 0,083 contre 0,167 pour les classes de très bon et mauvais état) et non spécifiques de ces

contextes (faible S : la médiane est comprise entre 0,36 et 0,52, alors qu'elle est supérieure à 0,88 pour les classes de très bon et mauvais état). La classe médiocre (4) peut avoir des motifs avec des valeurs plus élevées de P (médiane > 0,5) pour des longueurs de séquences de 12 et 18 mois. La classe très bon état (1) possède des motifs avec des valeurs de P élevées (médiane > 0.75) sauf pour la longueur de séquences la plus courte (3 mois). La classe de mauvais état possède les motifs aux valeurs de P les plus élevées (médiane >1,5) quelle que soit la longueur des séquences. Par indice biologique, les valeurs de P sont les plus dispersées avec des médianes supérieures à 0,5 pour l'IBD, l'IBMR, l'IBGN pour les longueurs de séquences de 6, 12 et 24 mois. A 3 mois, c'est le cas seulement pour l'IBGN. A 18 mois, c'est le cas pour l'IBD, l'IBMR et l'IPR. Quelle que soit la longueur de séquences, les valeurs de P sont les plus faibles pour l'I2M2.

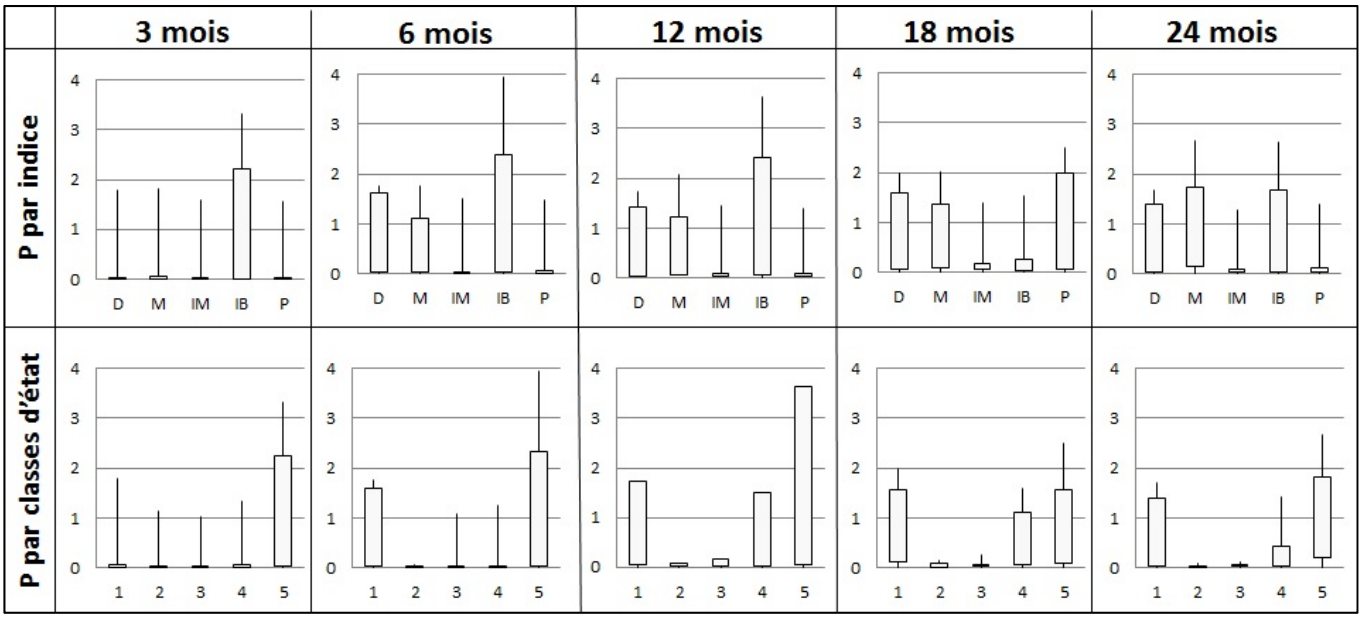


Figure 29 : Variation de la combinaison P des mesures d'intérêt calculées sur l'ensemble des motifs obtenus pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec le SEQ-eau , par indice biologique (D=IBD, M=IBMR, IM=I2M2, IB=IBGN, P=IPR) et par classe d'état (de 1 très bonne à 5 mauvaise) (les extrémités du rectangle représentent les 1ers et 3èmes quartiles, les extrémités des lignes les minima et maxima)

2.2 Caractérisation des altérations de l'ensemble des motifs obtenus par longueur de séquences

Les pourcentages d'apparition des altérations dans les données d'entrée et dans les motifs sont donnés dans le Tableau 8 et leur nombre représenté sur la Figure 30. Dans les données d'entrée, 59 combinaisons altérations et classes de qualité sont possibles. Nous en avons enlevé 10 avant d'extraire les motifs. Dans les motifs, ces combinaisons sont les items : leur nombre varie de 13 à 17 en fonction de la longueur des séquences. Pour 3 mois et 6 mois, ce nombre est respectivement de 16 et 17. Le maximum d'items est donc obtenu pour une fréquence minimale inférieure à 0,5 et une longueur de séquences de 6 mois. Les items qui n'apparaissent que pour ces deux extractions sont les matières phosphorées de qualité médiocre (PHOS orange), les matières organiques et oxydables de mauvaise qualité (MOOX rouge), les nitrates de qualité très bonne et médiocre (NITR bleu et NITR orange), les poly-chloro-biphényles de qualité moyenne (PCB jaune) à 3 mois, ainsi que l'effet prolifération végétale de bonne qualité (EPRV vert) à 6 mois. A 12, 18 et 24 mois, le nombre d'items est de 13. Les précédents ne sont plus présents, par contre les particules en suspension de qualité moyenne ou mauvaise (PAES jaune et rouge) apparaissent. Les autres items présents quelle que soit la longueur des séquences sont la minéralisation de bonne qualité (MINE vert), l'acidité de bonne qualité (ACID vert), les matières organiques de qualité très bonne et moyenne (MOOX bleu et jaune), les matières azotées de qualité très bonne et moyenne (AZOT bleu et jaune), les nitrates de qualité moyenne (NITR jaune), les matières phosphorées de qualité très bonne et moyenne (PHOS bleu et jaune), les pesticides de qualité médiocre ou mauvaise (PEST orange et rouge), les micropolluants hors pesticides de qualité moyenne et médiocre (MPOR jaune et orange), les micropolluants minéraux de qualité moyenne (MPMI jaune). Les items les plus nombreux dans les motifs sont les pesticides, les micropolluants minéraux hors pesticides, les nitrates, les matières organiques et l'acidité quelles que soient les longueurs de séquences, mais leur ordre varie : les pesticides sont toujours les plus nombreux (de 23,7 à 44,4%) excepté à 18 mois, où ce sont les nitrates les plus abondants (27,5%).

Tableau 8 : Pourcentages d'altérations présents dans les données d'entrée et dans l'ensemble des motifs obtenus pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec le SEQ-eau

Altérations	Données d'entrée	Motifs				
		3 mois	6 mois	12 mois	18 mois	24 mois
TEMP	8,9	0,0	0,0	0,0	0,0	0,0
ACID	8,8	7,5	11,5	9,1	16,8	6,9
MINE	8,6	0,0	2,0	1,1	1,2	1,0
PAES	8,1	0,0	0,0	0,0	2,0	3,8
EPRV	8,9	0,0	0,3	0,0	0,0	0,0
MOOX	8,9	5,1	5,8	9,1	13,1	10,7
AZOT	7,9	1,4	0,8	0,5	1,2	0,3
NITR	7,9	8,1	17,3	21,2	27,5	13,2
PHOS	7,9	2,7	2,3	1,3	0,8	1,0
PEST	5,8	44,4	25,8	23,7	18,4	40,7
MPOR	5,5	29,2	21,5	23,1	16,4	21,9
HAP	4,7	0,0	0,0	0,0	0,0	0,0
PCB	3,2	1,7	0,0	0,0	0,0	0,0
MPMI	4,8	0,0	13,0	10,8	2,5	0,5

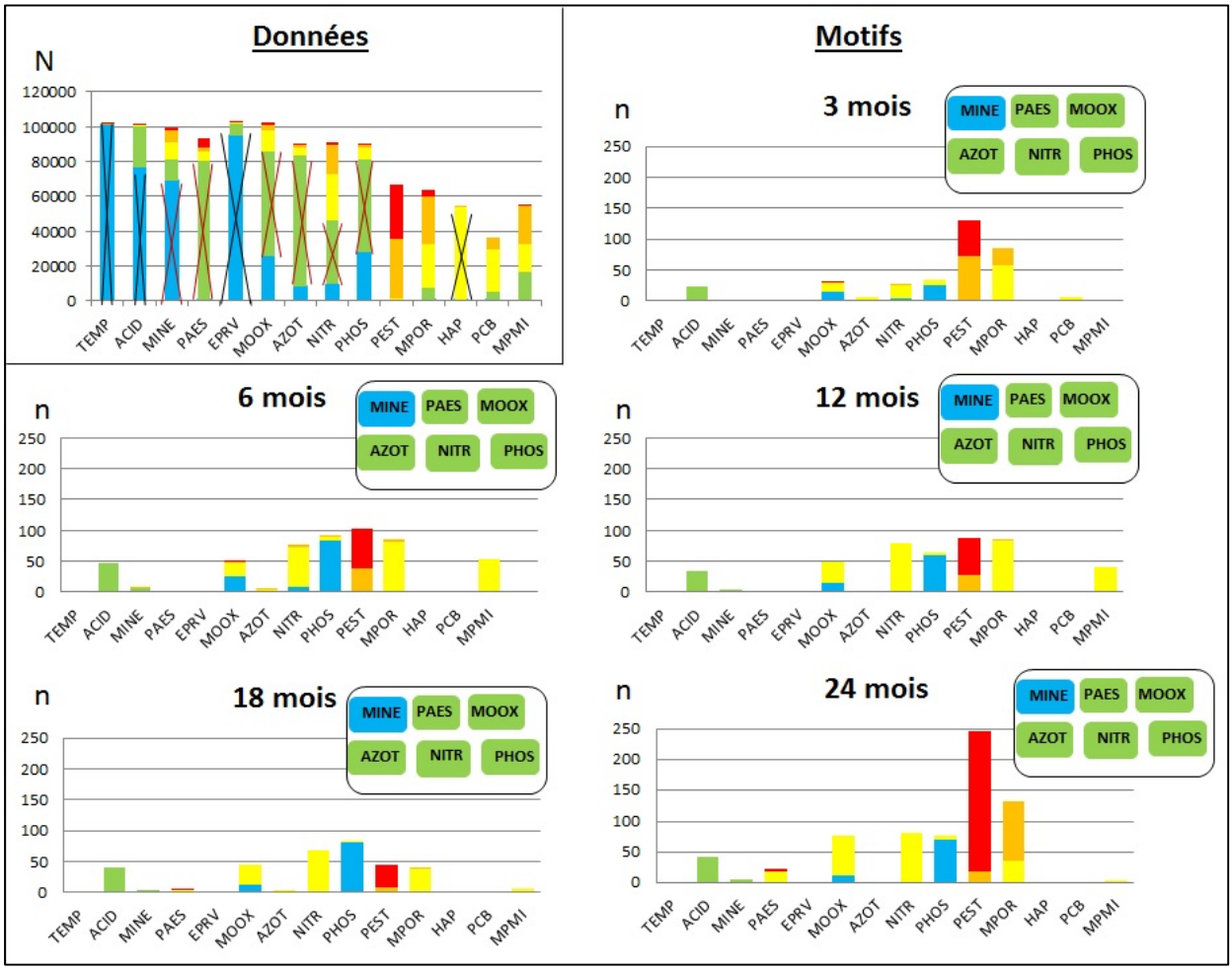


Figure 30 : Nombre d'altérations, par classe de qualité, présentes dans les données d'entrée et dans l'ensemble des motifs obtenus pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec le SEQ-eau (les altérations d'entrée barrées sont celles retirées avant extraction des motifs car trop abondantes dans les données d'entrée – en noir – ou trop abondantes dans les premières extractions de motifs – en rouge)

Pour chaque extraction, nous avons comptabilisé les altérations présentes dans l'ensemble des motifs par contexte. Les représentations complètes sous forme d'histogrammes sont données en annexe 3. Nous avons sélectionné ici les plus caractéristiques. Il est à noter que le nombre d'items est représenté en valeur absolue, aussi lorsque le nombre de motifs extraits pour un contexte donné est particulièrement élevé, le nombre d'items composant ces motifs est également important. C'est pour cette raison que le nombre des items apparaissant dans les motifs extraits pour le contexte IBMR rouge, à 24 mois sont systématiquement élevés. En effet, 74 motifs ont été extraits pour ce contexte contre 10 en moyenne sur les autres contextes, quelles que soit la longueur de séquences.

Parmi les macro-polluants, l'altération minéralisation (MINE) n'apparait dans les motifs que de bonne qualité, sa très bonne qualité n'ayant pas été conservée pour les extractions de motifs car dominantes dans les données d'entrée. L'item MINE_Vert n'apparait que pour les motifs extraits pour les contextes IBMR en mauvais état à partir de 12 mois et pour les contextes de l'IBD allant de moyen à mauvais état à partir de 6 mois. L'item EPRV_Vert n'apparait presque pas dans les motifs : il n'est caractéristique ni d'un indice, ni d'une longueur de séquences.

L'altération acidité (ACID) est uniquement observée dans les motifs en bonne qualité. Cet item apparait discriminant pour l'indice IBD (Figure 31) : le nombre d'items est toujours plus important pour l'IBD en très bon état (IBD bleu), quelle que soit la longueur de séquences. Pour les autres indices biologiques, le nombre de cet item est faible, le plus souvent nul ou égal à 1, quel que soit l'état de l'indice biologique, excepté pour le contexte IBMR rouge à 24 mois, mais qui est lié au nombre important de motifs pour ce contexte comme expliqué ci-dessus (annexe 3).

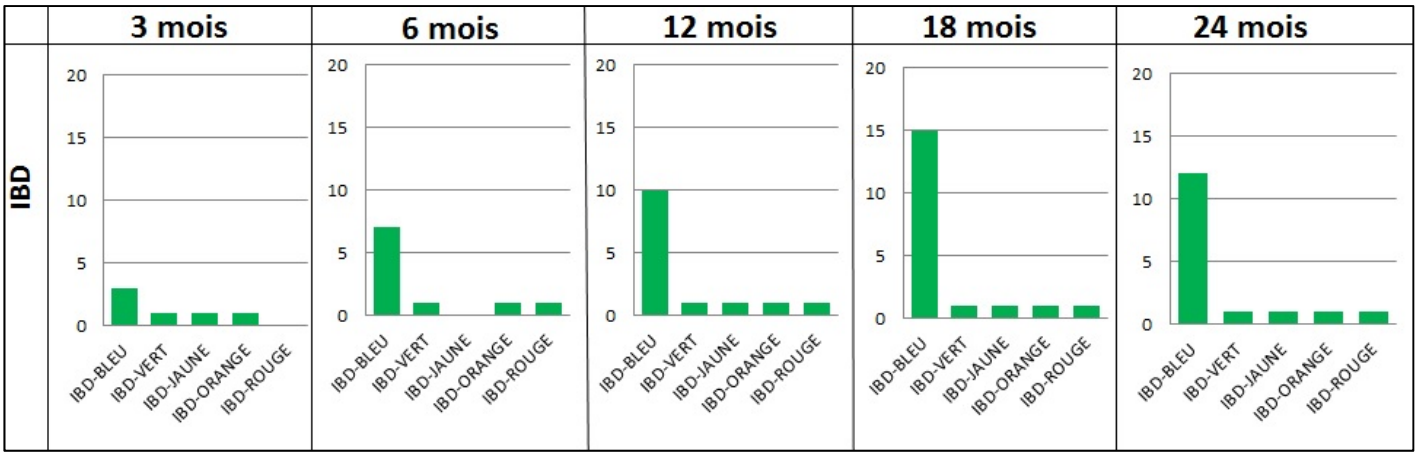


Figure 31 : Nombre d'apparitions des items de l'altération acidité (ACID) dans l'ensemble des motifs extraits pour les contextes de l'IBD par longueur de séquences 3, 6, 12, 18 et 24 mois, avec le SEQ-eau (à noter : extraction faite sans ACID_Bleu, dominante dans les données d'entrée)

Les items qui apparaissent dans les motifs pour l'altération particules en suspension (PAES) sont de qualité moyenne et mauvaise, la bonne qualité (PAES_Vert) ayant été retirée avant extraction car initialement dominante dans les premières extractions de motifs. Cet item apparait discriminant pour l'indice IBMR (Figure 32) essentiellement en mauvais état pour des longueurs de séquences supérieures ou égales à 12 mois. Cet item est aussi observé dans le contexte I2M2 en état mauvais à partir des séquences 6, 18 et 24 mois, mais à raison d'un seul item par contexte. Il n'apparait pas discriminant pour les contextes de l'IBGN (apparition d'un item pour l'IBGN bleu à 6 mois et un pour l'IBGN rouge à 18 mois). Il n'apparait pas dans les motifs extraits pour les indices IPR et IBD (cf Annexe 3).

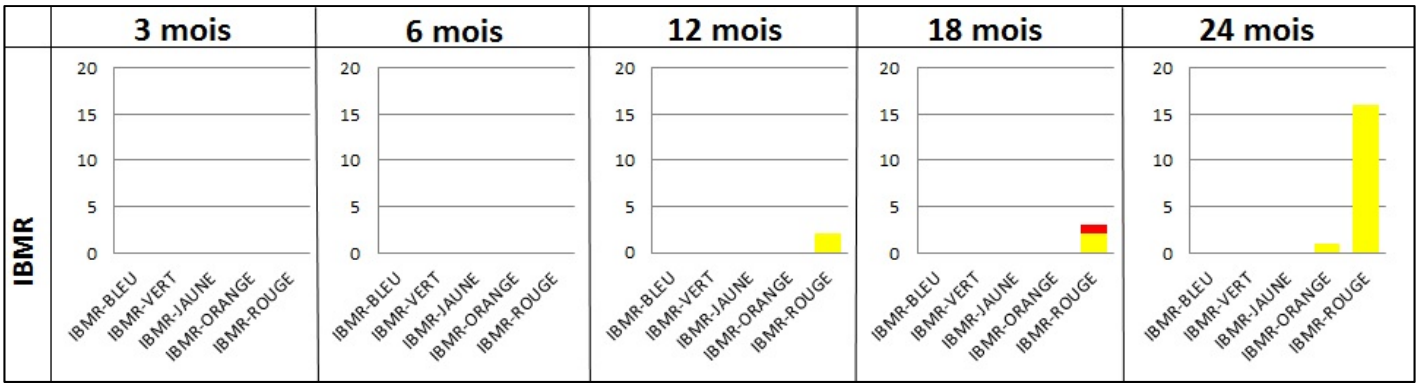


Figure 32 : Nombre d'apparitions des items de l'altération particules en suspension (PAES) extraits pour les contextes de l'IBMR par longueur de séquences 3, 6, 12, 18 et 24 mois, avec le SEQ-eau (à noter : extraction faite sans PAES_Vert, initialement dominante dans les premières extractions)

L'altération matières organiques et oxydables (MOOX) est observée dans les motifs seulement de qualité très bonne, moyenne et , dans une moindre mesure, mauvaise, la bonne qualité (MOOX_Vert) ayant été retirée avant extraction car initialement dominante dans les premières extractions de motifs (Figure 33). L'item MOOX_rouge n'apparait que pour le contexte IBGN en mauvais état (IBGN_rouge) pour la longueur de séquences 3 mois, ceci étant sûrement dû à la faible fréquence minimale (0,30) utilisée pour cette extraction. Les deux autres items apparaissent quelles que soient les longueurs de séquences et quels que soient les indices. MOOX_bleu est trouvé dans les motifs extraits pour les contextes en très bon et bon état de chaque indice biologique, mais aussi pour les états moyens de l'IBMR à 3 et 6 mois et de l'IPR à 3 mois. MOOX_jaune est trouvé dans les motifs de tous les indices biologiques, dans les états médiocre et mauvais pour l'IBMR, l'IPR et l'I2M2, dans les états moyen et médiocre pour l'IBD, et pour les états moyens, médiocre et mauvais de l'IBGN quelle que soit la longueur des séquences.

L'altération matières azotées (AZOT) n'est observée presque que de qualité moyenne, la bonne qualité (AZOT_Vert) ayant été retirée avant extraction car initialement dominante dans les premières extractions de motifs. Elle n'apparait que pour les motifs des contextes de l'IBMR mauvais (1 à 3 items) et de l'I2M2 mauvais (un seul item) quelles que soient les longueurs de séquences ; à l'exception près d'un motif avec l'item AZOT_jaune pour l'IBGN en très bon état à 6 mois et d'un motif avec l'item AZOT_bleu pour l'IPR en très bon état à 3 mois (cf Annexe 3).

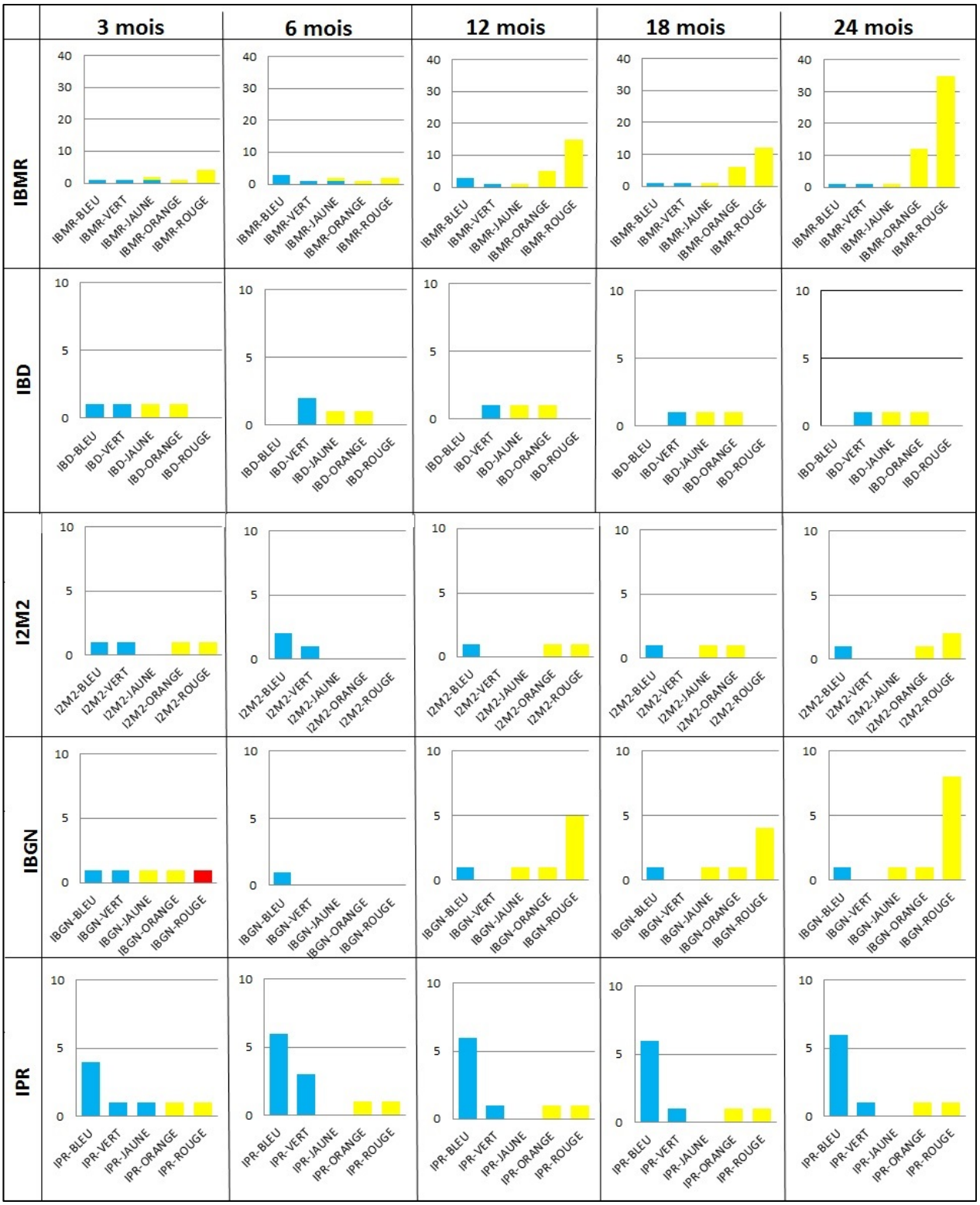


Figure 33 : Nombre d'apparitions des items de l'altération matières organiques et oxydables (MOOX) extraits pour l'ensemble des indices biologiques et par longueur de séquences 3, 6, 12, 18 et 24 mois, avec le SEQ-eau (à noter : extraction faite sans MOOX_Vert, initialement dominante dans les premières extractions)

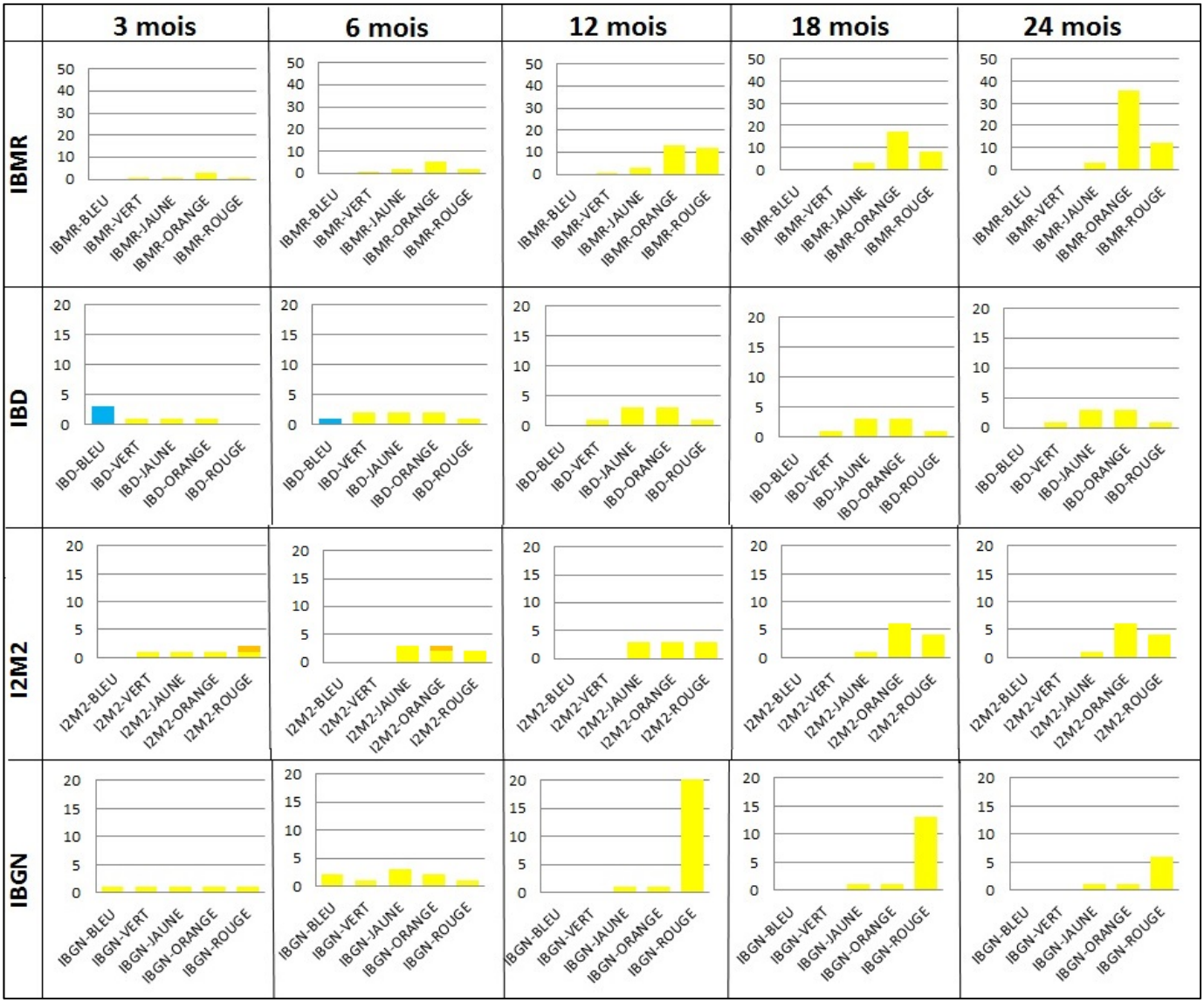


Figure 34 : Nombre d'apparitions des items de l'altération nitrates (NITR) extraits pour les contextes de l'IBMR, l'IBD et l'IBGN, par longueur de séquences 3, 6, 12, 18 et 24 mois, avec le SEQ-eau (à noter : extraction faite sans NITR_Vert, initialement dominante dans les premières extractions)

L'altération nitrates (NITR) est observée dans les motifs de qualité très bonne, moyenne et médiocre, la bonne qualité (NITR_Vert) ayant été retirée avant extraction car initialement dominante dans les premières extractions de motifs (Figure 34). Cette altération n'est pas discriminante dans les motifs extraits pour les contextes de l'IPR. L'item NITR_orange n'est observé que dans deux motifs extraits pour l'I2M2 en mauvais état à 3 mois et en état médiocre à 6 mois. L'item NITR_bleu n'est observé que dans 1 à 2 motifs extraits pour l'IBD en très bon état à 3 et 6 mois. Ces deux items n'apparaissent plus pour les motifs extraits pour ces deux indices lorsque les fréquences minimales sont supérieures à 0,4, à partir des longueurs de séquences de 12 mois. L'item NITR_jaune semble peu discriminant dans les motifs des contextes de l'IBD et de l'I2M2 : son nombre d'apparitions reste limité entre 1 et 3 dans l'ensemble des motifs pour l'IBD des états bon à mauvais, et entre 1 et 6 dans l'ensemble des motifs pour l'I2M2 des états moyen à mauvais. Par contre, pour des séquences au moins égales à 12 mois, le nombre d'apparitions de cet item augmente significativement dans les motifs extraits pour les contextes IBMR médiocre et mauvais et le contexte IBGN mauvais.

L'altération phosphates (PHOS) est observée dans les motifs seulement de qualité très bonne et moyenne, la bonne qualité (PHOS_Vert) ayant été retirée avant extraction car initialement dominante dans les premières extractions de motifs (Figure 35). Cette altération est peu discriminante dans les motifs extraits pour les contextes de l'IBGN et l'IPR, non représentés sur la Figure 35. L'item PHOS_jaune n'est observé que dans quelques motifs extraits pour l'I2M2 en état médiocre ou mauvais et l'IBMR en état mauvais quelles que soient les longueurs de séquences. L'item PHOS_bleu est observé pour les contextes IBMR et I2M2 en très bon et bon état quelles que soient les longueurs de séquences. Cet item est surtout très présent dans les motifs extraits pour le contexte IBD en très bon état : son nombre d'apparitions est croissant avec la longueur des séquences.

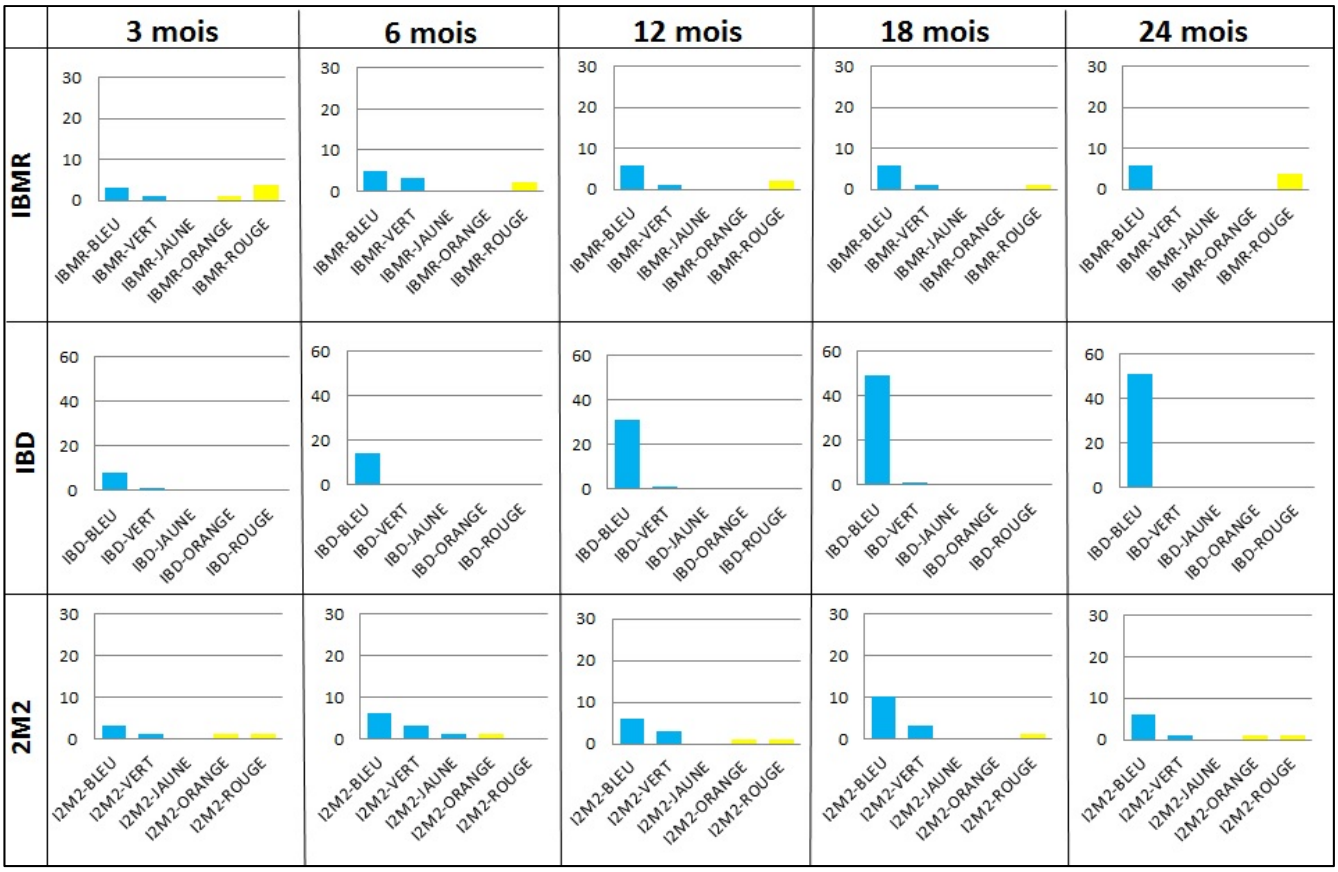


Figure 35 : Nombre d'apparitions des items de l'altération phosphates (PHOS) extraits pour les contextes de l'IBMR, l'IBD et l'I2M2, par longueur de séquences 3, 6, 12, 18 et 24 mois, avec le SEQ-eau (à noter : extraction faite sans PHOS_Vert, initialement dominante dans les premières extractions)

Parmi les micropolluants, l'altération pesticides (PEST) est présente dans les motifs extraits dans les qualités médiocre et mauvaise. Ces deux items apparaissent dans quasiment tous les contextes de tous les indices quelle que soit la longueur des séquences. Bien que très présents, et en nombre particulièrement important pour l'IBMR en états médiocre et mauvais à 24 mois (respectivement 24 et 144 apparitions), ces items semblent peu discriminants que ce soit par indice ou par longueurs de séquences (annexe 3).

L'altération des poly-chloro-biphényles (PCB) n'apparait que de qualité moyenne (PCB_jaune) uniquement dans les motifs extraits à 3 mois pour les contextes I2M2 en mauvais état, IBGN en états médiocre et mauvais et l'IPR en état

très bon et mauvais. La présence de cet item dans les motifs ne semble possible que pour les fréquences minimales inférieures ou égales à 0,3.

L'altération micropolluants organiques hors pesticides (MPOR) n'apparait dans les motifs que de qualité moyenne et médiocre. Pour l'IBMR, cet item est essentiellement présent, de qualité médiocre, dans les motifs extraits pour le contexte mauvais, à 24 mois (92 apparitions). Pour les contextes de l'IBD, l'I2M2 et l'IPR, ces items n'apparaissent presque que pour la longueur de séquences 3 mois, mais sont peu discriminants d'un état. Par contre, pour l'IBGN, ces items apparaissent essentiellement dans la qualité moyenne pour le contexte état mauvais et quelle que soit la longueur de séquences.

L'altération micropolluants minéraux (MPMI) n'apparait que dans la qualité médiocre et seulement dans les motifs extraits pour l'IBGN, dans son contexte mauvais (Figure 36), principalement pour les longueurs de séquences 3, 12 mois et 6mois.

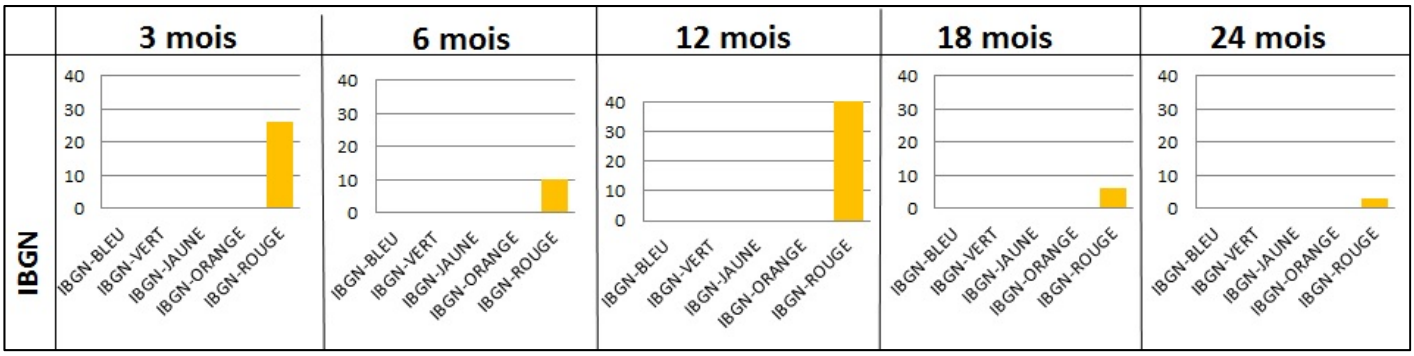


Figure 36 : Nombre d'apparitions des items de l'altération micropolluants minéraux (MPMI) dans l'ensemble des motifs par altérations, pour l'IBGN par longueur de séquences 3, 6, 12, 18 et 24 mois, avec le SEQ-eau

2.3 Motifs caractéristiques obtenus par longueurs de séquences

Les cinq motifs caractéristiques ont été sélectionnés par contexte sur la base de leur combinaison P la plus élevée, ainsi que de leurs mesures d'intérêt et items. Ils sont disponibles en annexes 4a à 4e par longueur de séquences de 3 à 24 mois.

Parmi ces cinq motifs sélectionnés, nous n'avons représenté ci-après que les motifs des contextes ayant des motifs dominants. Ces contextes sont l'IBD et l'IPR en très bon état (Figure 37 et Figure 38), l'IBMR en états médiocre et mauvais (Figure 39 et Figure 40), et l'I2M2 et l'IBGN en mauvais état (Figure 41 et Figure 42).

Pour le contexte IBD bleu (Figure 37) sont observés des motifs dominants pour chaque longueur de séquences. Les items associés sont essentiellement l'acidité de bonne qualité (ACID_Vert) et les matières phosphorées de très bonne qualité à chaque longueur de séquence. L'item des nitrates de très bonne qualité est associé à l'item PHOS_Bleu uniquement pour 3 et 6 mois (respectivement motifs n° 59 et 71). Le nombre d'items PHOS_Bleu se succédant augmente avec les longueurs de séquences de 1 à 3 mois jusqu'à 7 à 24 mois (motif n° 58). Ce sont les seuls items qui apparaissent dans les cinq motifs sélectionnés pour ce contexte quelle que soit la longueur des séquences.

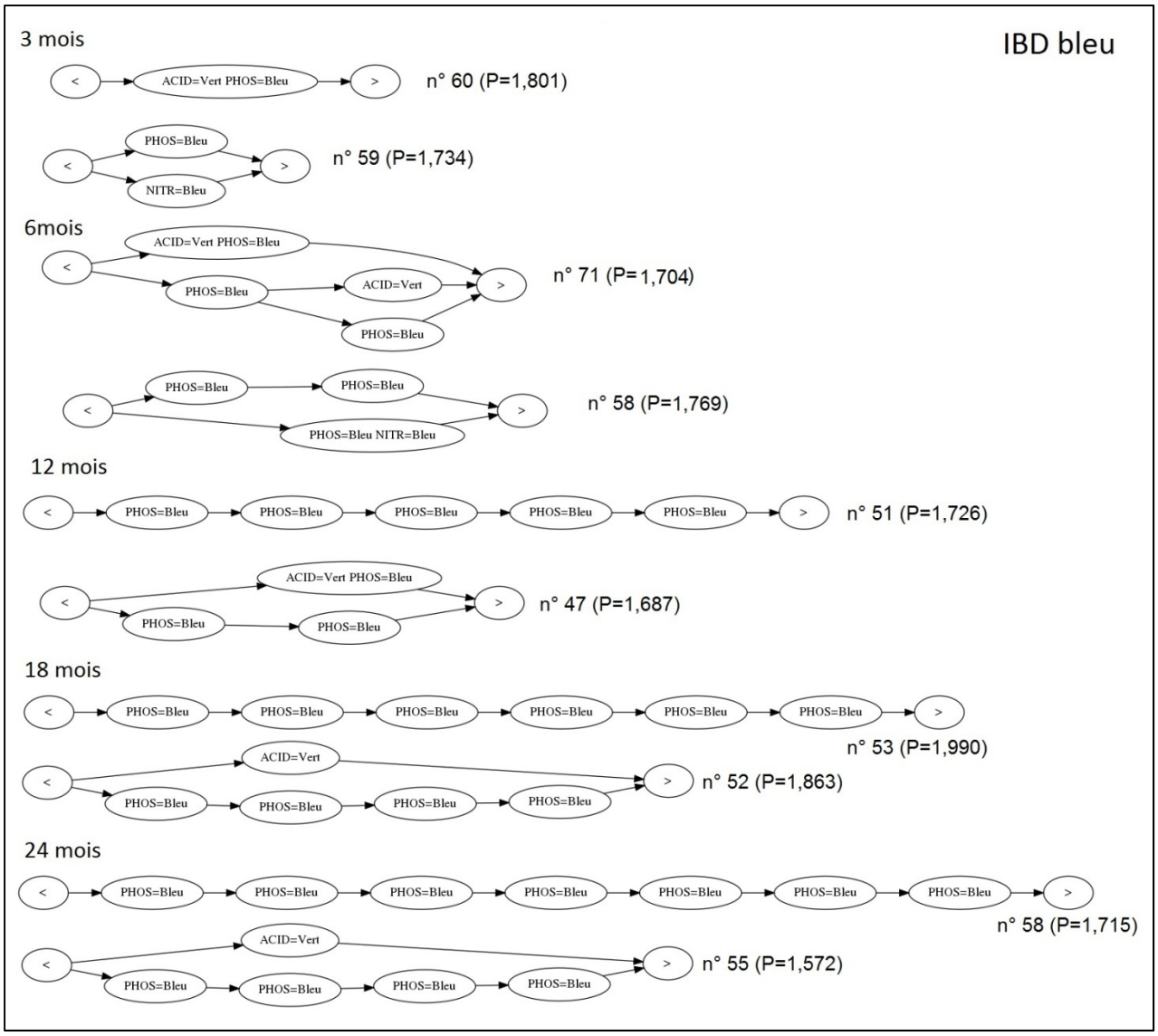


Figure 37 : Motifs extraits pour le contexte IBD en très bon état, choisis parmi les 5 motifs sélectionnés sur la base de leur combinaison P (indiquée entre parenthèses) la plus élevée, pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec le SEQ-eau

Les motifs caractéristiques de l'IPR en très bon état (Figure 38) se limitent principalement à des items des matières organiques et oxydables (MOOX_Bleu) se répétant 2 fois à 3 fois, quelles que soient les longueurs de séquences. L'item des matières azotées hors nitrates en très bon état (AZOT_Bleu) n'apparait que dans un seul motif à 3 mois (n° 216). C'est le cas également de l'item matières organiques et oxydables (MOOX_Bleu) qui apparait une seule fois associé à l'item PHOS_Bleu à 6 mois (motif n° 244). Il est à noter qu'à 12 mois, parmi les cinq motifs sélectionnés

pour ce contexte, le motif n° 225 est dominant et composé d'un item PHOS_bleu associé à un item pesticides de qualité médiocre (PEST_Orange) (Annexe 4c).

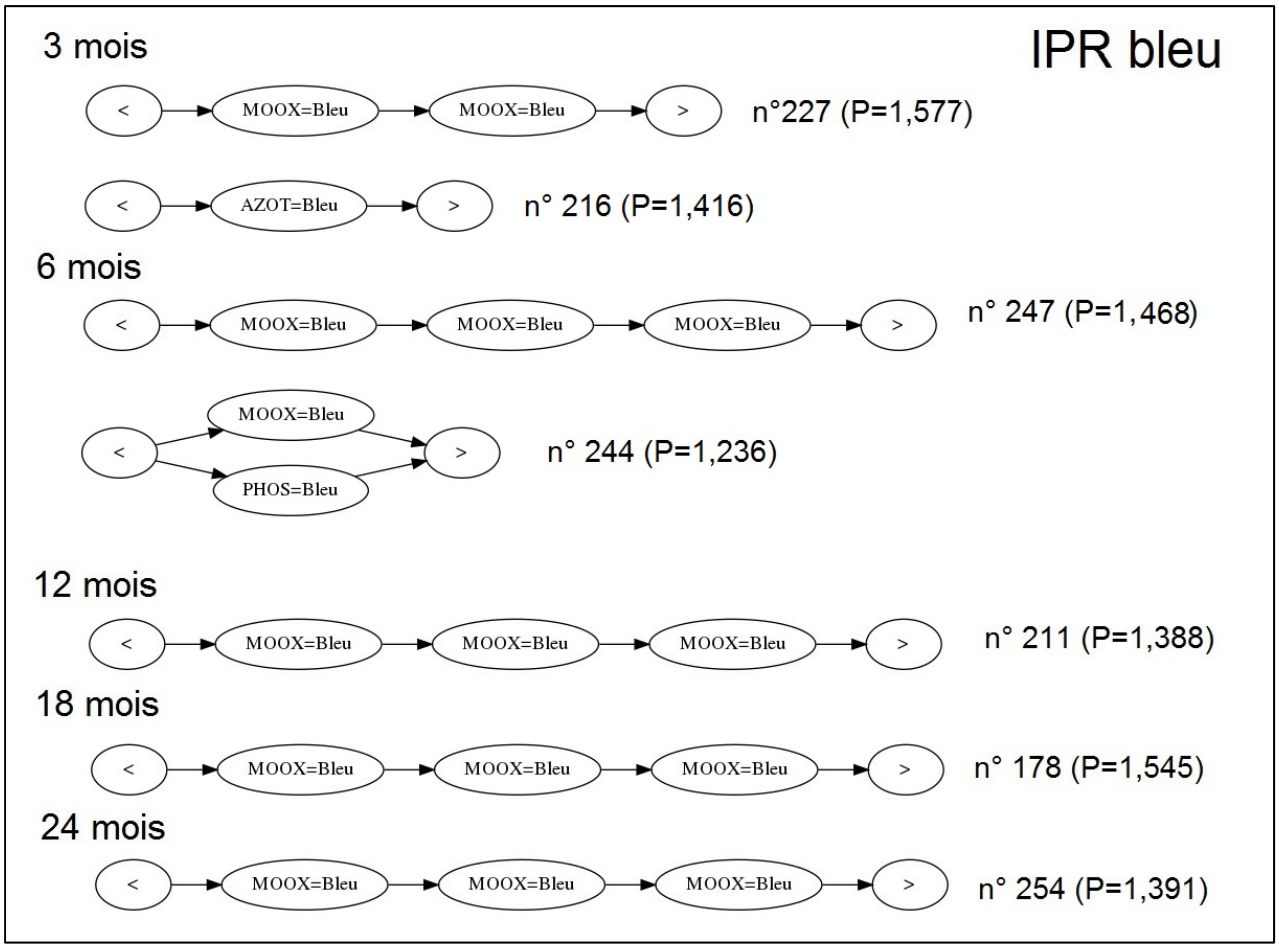


Figure 38 : Motifs extraits pour le contexte IPR en très bon état, choisis parmi les 5 motifs sélectionnés sur la base de leur combinaison P (indiquée entre parenthèses) la plus élevée, pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec le SEQ-eau

Les motifs caractéristiques du contexte IBMR en état médiocre (Figure 39) sont principalement composés de l'item nitrates de qualité moyenne (NITR_Jaune) se répétant de 2 à 4 fois pour 3 à 18 mois. Cet item peut être associé à l'item de l'acidité en bon état (ACID_Vert, motif n° 208) à 6 mois ou à l'item des matières organiques de qualité moyenne (MOOX_Jaune, motif n° 162) à 24 mois. Enfin, parmi les motifs dominants sélectionnés pour ce contexte, l'item pesticides de mauvaise qualité (PEST_Rouge) apparait à 12, 18 et 24 mois (respectivement motifs n° 175, 145 et 154, voir annexes 4 c à 4e) où il est associé à un ou deux items NITR_Jaune (motifs n° 175 et 154) ou seul (motif n° 145 à 4 items PEST_Rouge).

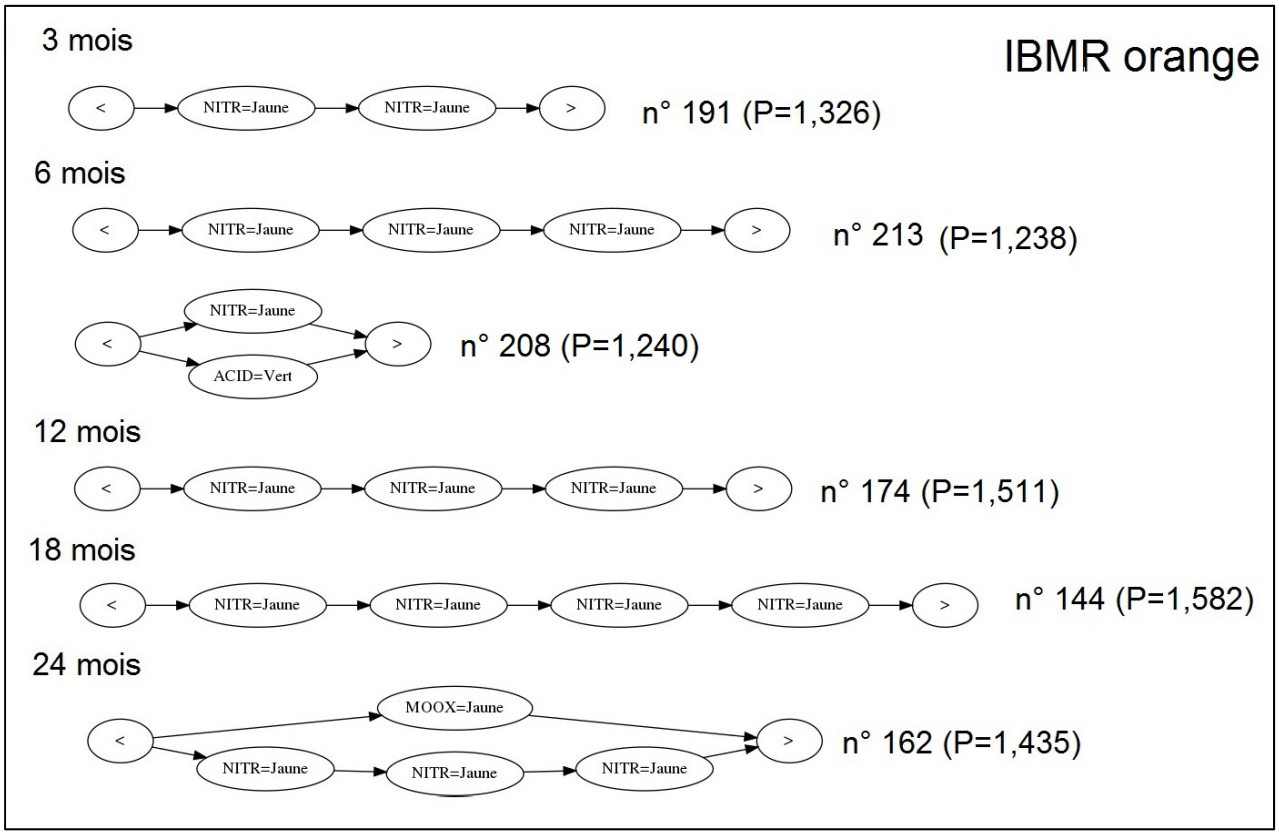


Figure 39 : Motifs extraits pour le contexte IBMR en état médiocre, choisis parmi les 5 motifs sélectionnés sur la base de leur combinaison P (indiquée entre parenthèses) la plus élevée, pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec le SEQ-eau

Les motifs caractéristiques du contexte IBMR en mauvais état (Figure 40) sont composés de plusieurs items : les matières organiques de qualité moyenne (MOOX_Jaune) sont présentes dans tous les motifs retenus de 3 à 24 mois. Ces items peuvent être associés à l'item des nitrates de qualité moyenne (NITR_Jaune, motifs n° 229 à 6 mois, n° 197 à 12 mois), à l'item des matières azotées hors nitrates de qualité moyenne (AZOT_jaune, motif n° 202, à 3 mois), ou aux matières phosphorées de qualité moyenne (PHOS_Jaune, motif n° 204 à 3 mois). A 24 mois, tous les motifs sélectionnés pour ce contexte contiennent également des items des micropolluants organiques hors pesticides de qualité médiocre (MPOR_Orange) et

des pesticides de mauvaise qualité (PEST_Rouge) (motifs n° 217, mais aussi n° 211, 203, 199 et 215, annexe 4e). A partir de 12 mois, les items des particules en suspension de qualité moyenne et mauvaise (PAES_Jaune et PAES_Rouge) sont présents parmi les motifs sélectionnés (motifs n° 191 et 190, à 12 mois, motifs n° 156 et 153 à 18 mois et n° 211 et 203 à 24 mois, annexe 4e).

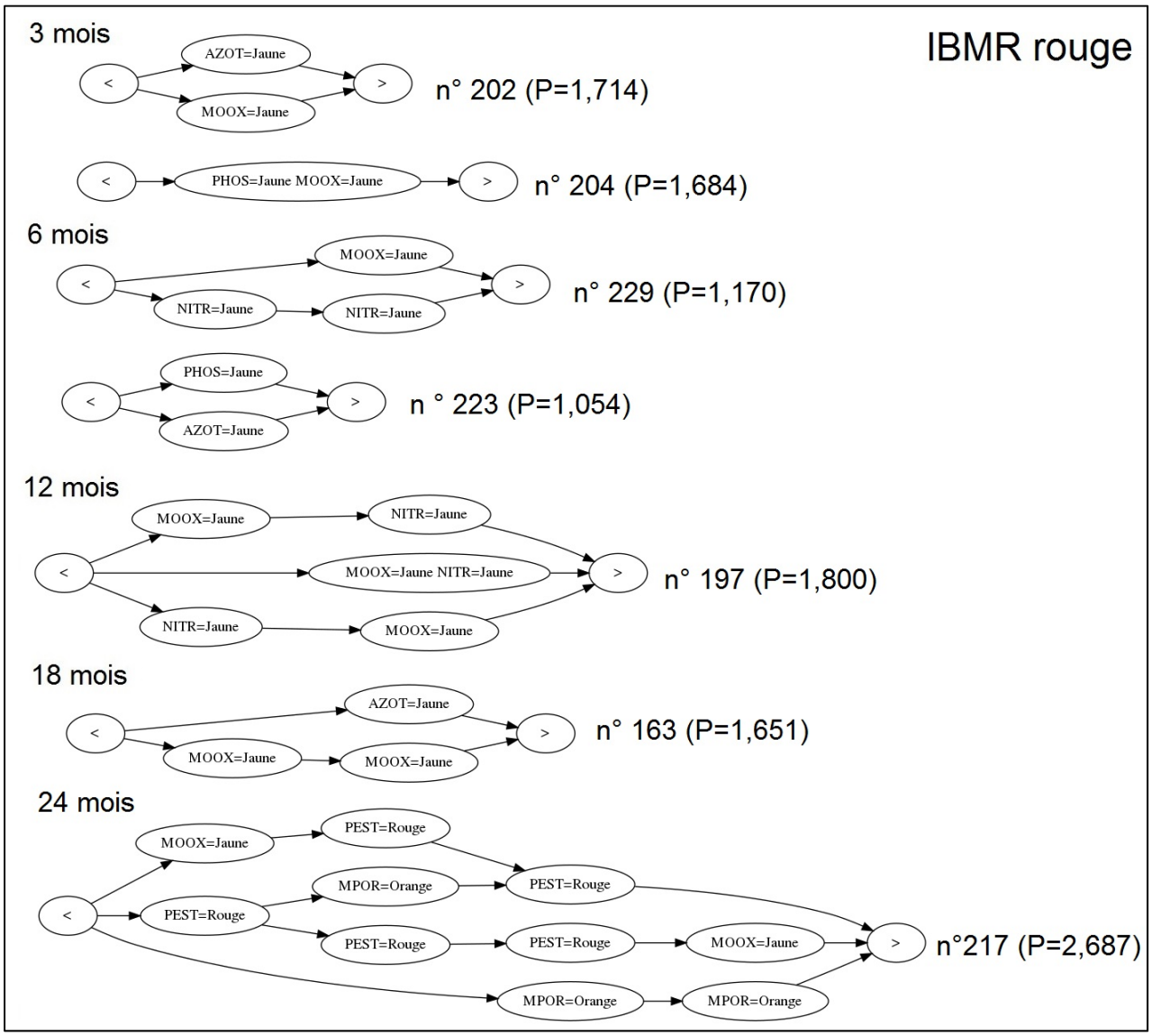


Figure 40 : Motifs extraits pour le contexte IBMR en mauvais état, choisis parmi les 5 motifs sélectionnés sur la base de leur combinaison P (indiquée entre parenthèses) la plus élevée, pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec le SEQ-eau

Les motifs caractéristiques du contexte IBGN en mauvais état sont complexes dès 3 mois : le motif n°150 a 5 items et 4 itemsets (Figure 41). Les items présents parmi les macro-polluants sont les nitrates de qualité moyenne (NITR_Jaune, motifs n° 170 à 6 mois, 120 à 12 mois, 104 à 18 mois et 112 à 24 mois), et les matières organiques de qualité moyenne (MOOX_Jaune, motif n° 112 à 24 mois). Tous les motifs de ce contexte contiennent des items de micropolluants : micropolluants organiques hors pesticides de qualité moyenne (MPOR_Jaune) et des micropolluants minéraux de qualité médiocre (MPMI_Orange) quelle que soit la longueur des séquences. L'item des pesticides de qualité médiocre (PEST_Orange) est également présent dans une partie des motifs sélectionnés à 3 mois (motifs n° 150, 155 et 145, annexe 4a) et 6 mois (motifs n° 169 et 167, annexe 4b).

Les motifs caractéristiques du contexte I2M2 en mauvais état sont limités à un seul item quelle que soit la longueur des séquences (Figure 42). L'item des matières azotées hors nitrates de qualité moyenne (AZOT_Jaune) est le plus fréquent (motifs 26 à 6 mois, 23 à 12 mois, n° 24 à 24 mois). Sont présents également l'item des matières phosphorées de qualité moyenne (PHOS_Jaune, motifs n° 30 à 6 mois et 20 à 18 mois) et deux autres items rares pour les autres contextes : les PCB de qualité moyenne (PCB_Jaune, motif n° 29 à 3 mois) et les nitrates de qualité médiocre (NITR_Orange, motif n° 36 à 3 mois et 28 à 6 mois). Les nitrates de qualité médiocre étaient présents dans les données d'entrée, mais nous avons trouvé très peu d'items leur correspondant dans les motifs, mis à part pour l'I2M2.

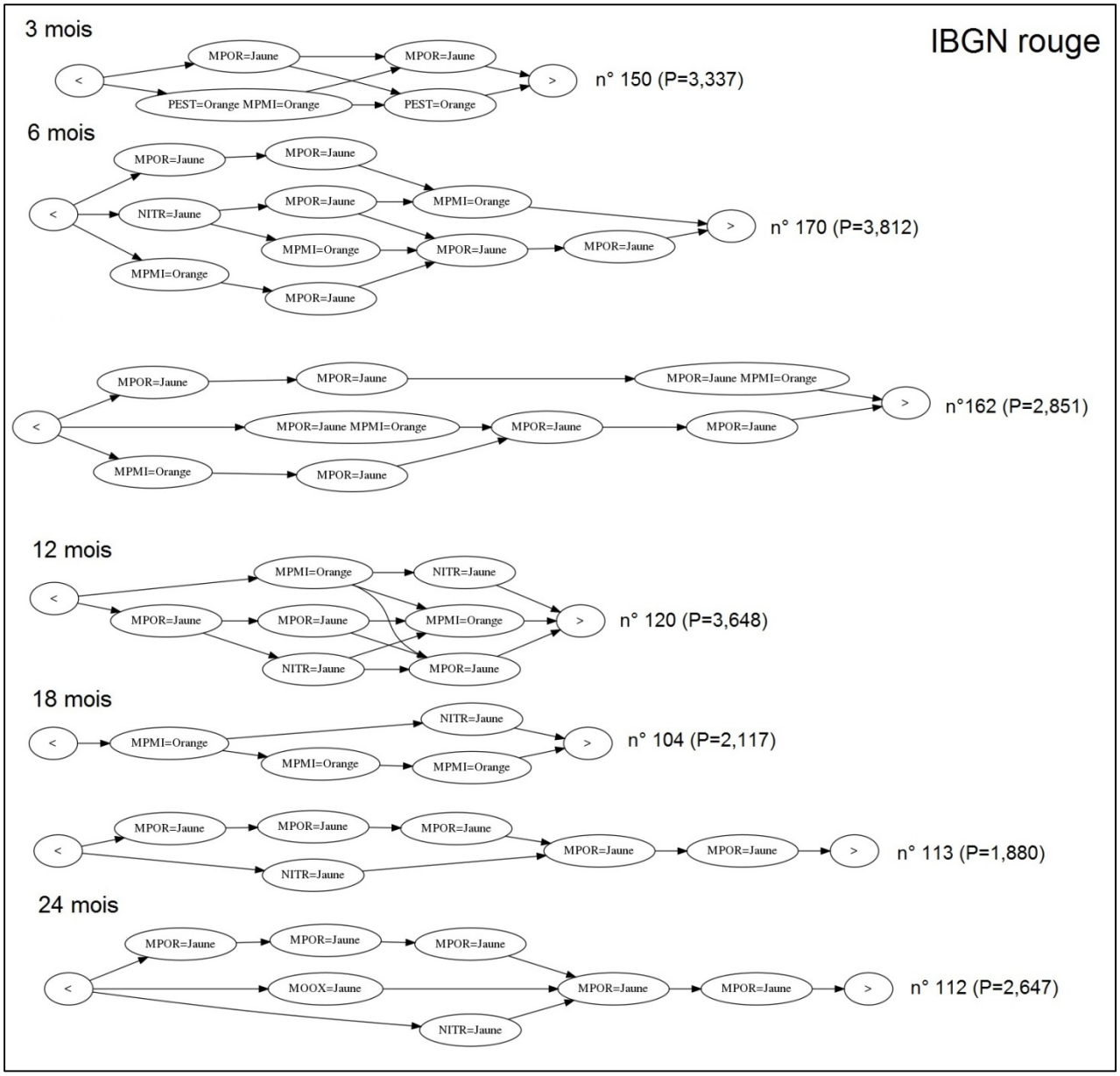


Figure 41 : Motifs extraits pour le contexte IBGN en mauvais état, choisis parmi les 5 motifs sélectionnés sur la base de leur combinaison P (indiquée entre parenthèses) la plus élevée, pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec le SEQ-eau

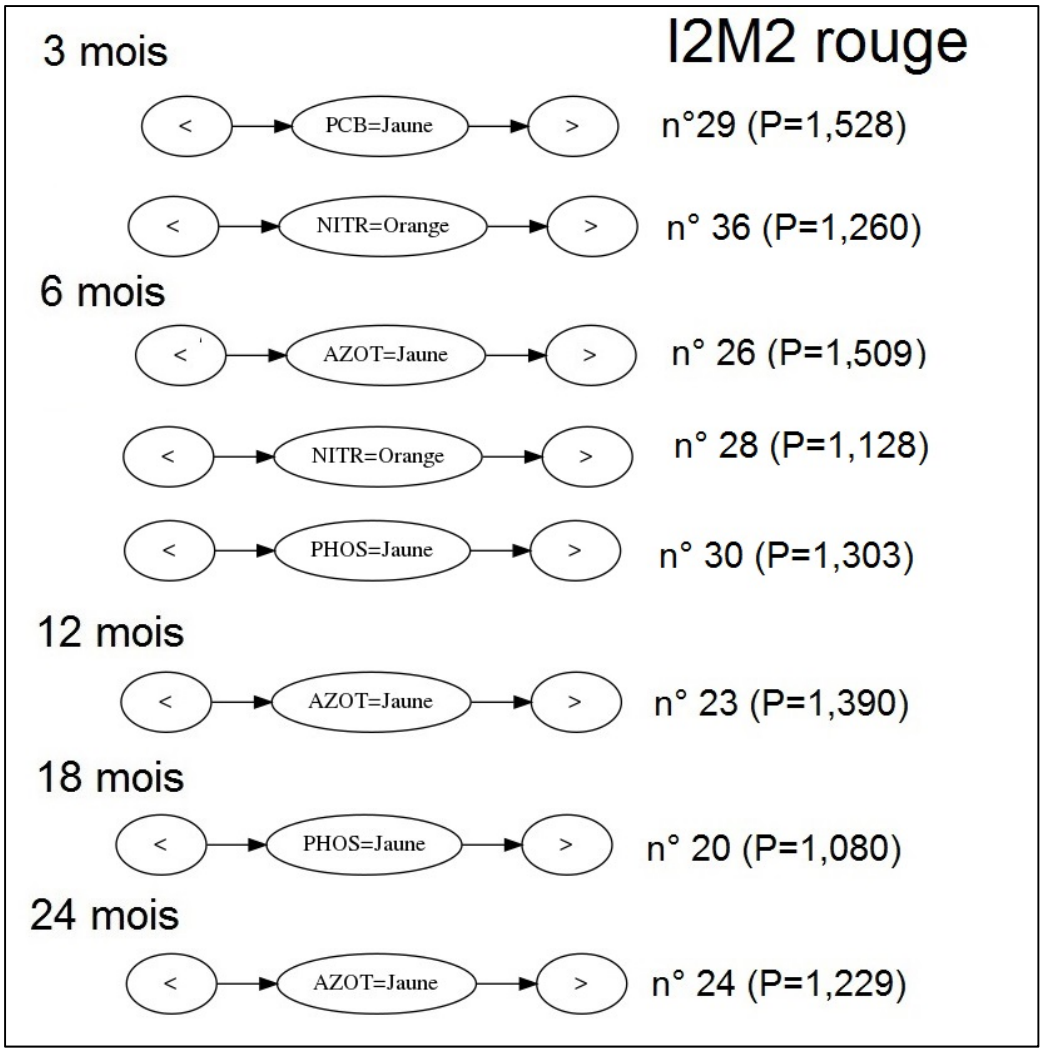


Figure 42 : Motifs extraits pour le contexte I2M2 en mauvais état, choisis parmi les 5 motifs sélectionnés sur la base de leur combinaison P (indiquée entre parenthèses) la plus élevée, pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec le SEQ-eau

3 Résultats obtenus pour les grilles DCE

3.1 Caractérisation des ensembles de motifs obtenus par longueurs de séquences

Le nombre de motifs obtenus avec les grilles DCE est plus faible mais plus équilibré par contexte, qu'avec la grille SEQ-eau : le maximum de nombres de motifs obtenus est de 40 pour le contexte de l'IBMR d'état médiocre pour la longueur de séquences de 12 mois contre 74 motifs extraits pour l'IBMR en mauvais état, avec le

SEQ-eau, à 24 mois. Comme avec le SEQ-eau, un nombre plus important de motifs, dont des motifs dominants est trouvé pour les contextes des classes extrêmes de l'IBD en très bon état et de l'IBMR états médiocre et mauvais et de l'IBGN en mauvais état. Mais, contrairement à ce qui a été observé avec le SEQ-eau, c'est aussi le cas pour l'IBD en bon état, l'IPR et l'IBMR en très bon état.

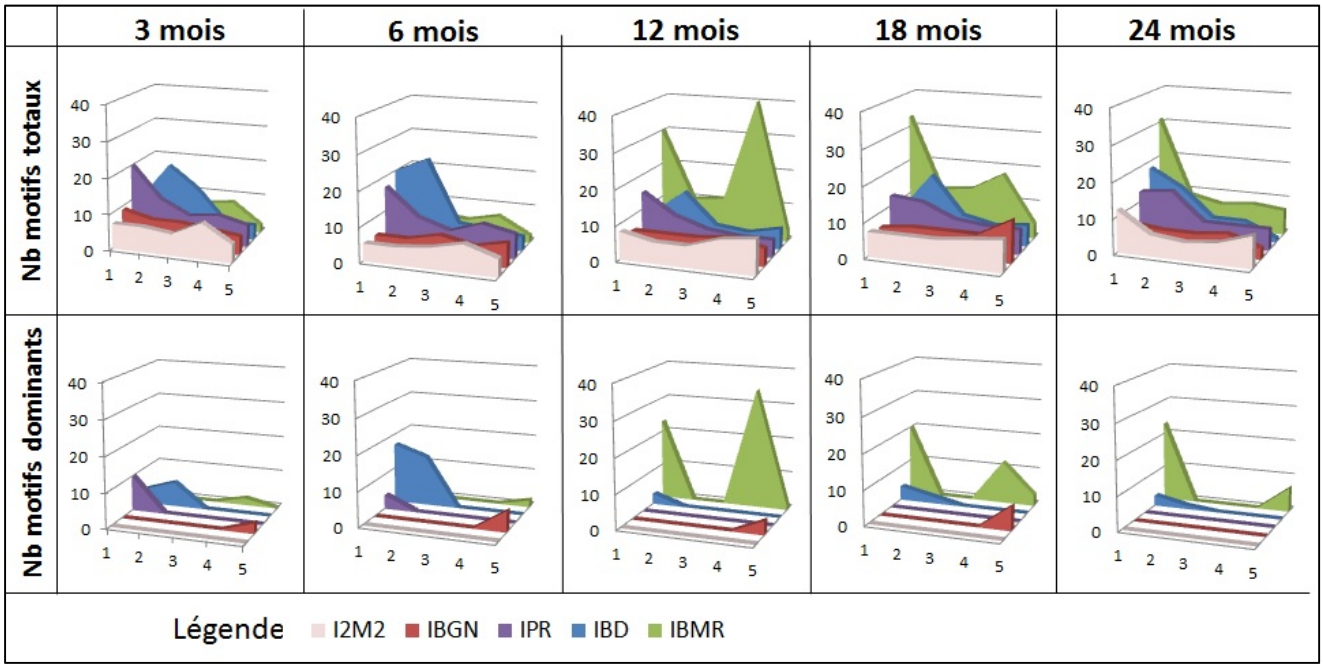


Figure 43 : Nombre de motifs et de motifs dominants par indice biologique et par classes des grilles DCE (de 1, très bonne à 5 mauvaise) obtenu pour les extractions réalisées pour les longueurs de séquences 3, 6, 12, 18 et 24 mois

Les motifs extraits avec les grilles DCE sont en moyenne de la même longueur que ceux obtenus avec le SEQ-eau : de 2 items pour les longueurs de séquences 3, 6 et 12 mois à 3 items pour 18 et 24 mois. Les motifs sont plus courts que ceux obtenus avec le SEQ-eau : le maximum est de 9 items, pour l'extraction faite pour la longueur de séquences 24 mois, contre 12 (longueurs de séquences 6 et 24 mois, avec le SEQ-eau).

Le Tableau 9 fournit les principales valeurs de dispersion – médiane, minimum, maximum, 1er et 3ème quartiles et variance – des mesures d'intérêt F, C, S, E et de leur combinaison P calculées sur l'ensemble des motifs obtenus pour les extractions réalisées pour les cinq longueurs de séquences de 3, 6, 12, 18 et 24

mois. Sur la Figure 44 sont représentés les quartiles, les minima et maxima en fonction des cinq longueurs de séquences, par mesures d'intérêt et combinaisons. Il y a moins de motifs dominants (émergence E ≠ 0) obtenus avec les grilles DCE : le maximum de l'émergence E varie de 1,113 à 1,458 contre 2,117 à 3,200 avec le SEQ-eau. Les motifs sont également moins singuliers – ils sont plus souvent partagés par de nombreux contextes – : la médiane de S est comprise entre 0,083 et 0,880 contre 0,160 à 0,520 pour les extractions avec le SEQ-eau.

Tableau 9 : Variations –médiane, variance, écart-type, 1er et 3ème quartile, minimum, maximum, - des mesures d'intérêt F, C, S, E et de leur combinaison P calculées sur l'ensemble des motifs obtenus pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec les grilles DCE

Mesures d'intérêt	Longueur séquences (mois)	MEDIANE	VARIANCE	ECART-TYPE	1ère QUARTILE	3ème QUARTILE	MIN	MAX
F	3	0,559	0,006	0,079	0,510	0,608	0,480	0,864
	6	0,658	0,006	0,080	0,623	0,724	0,600	0,945
	12	0,799	0,002	0,047	0,772	0,841	0,760	0,968
	18	0,837	0,002	0,041	0,812	0,867	0,800	0,974
	24	0,910	0,001	0,024	0,892	0,927	0,880	1,000
C	3	0,333	0,021	0,155	0,167	0,333	0,024	1,000
	6	0,333	0,028	0,168	0,167	0,333	0,167	1,000
	12	0,286	0,035	0,187	0,143	0,429	0,143	1,000
	18	0,333	0,067	0,259	0,167	0,667	0,167	1,000
	24	0,333	0,030	0,174	0,222	0,444	0,111	1,000
S	3	0,160	0,115	0,340	0,080	0,720	0,040	0,960
	6	0,360	0,124	0,352	0,120	0,800	0,040	0,960
	12	0,520	0,141	0,377	0,080	0,880	0,040	0,960
	18	0,280	0,109	0,331	0,200	0,820	0,040	0,960
	24	0,520	0,111	0,333	0,200	0,800	0,040	0,960
E	3	1,072	0,017	0,130	1,034	1,105	1,002	1,458
	6	1,108	0,023	0,151	1,050	1,337	1,005	1,421
	12	1,053	0,002	0,047	1,035	1,077	1,001	1,275
	18	1,024	0,001	0,028	1,016	1,054	1,003	1,113
	24	1,042	0,002	0,047	1,029	1,054	1,003	1,279
P=FCS+E	3	0,020	0,178	0,422	0,008	0,131	0,008	1,636
	6	0,081	0,360	0,600	0,015	0,244	0,004	2,025
	12	0,129	0,363	0,604	0,019	1,177	0,004	1,777
	18	0,152	0,303	0,552	0,035	0,473	0,005	1,845
	24	0,062	0,267	0,518	0,029	0,261	0,016	2,148

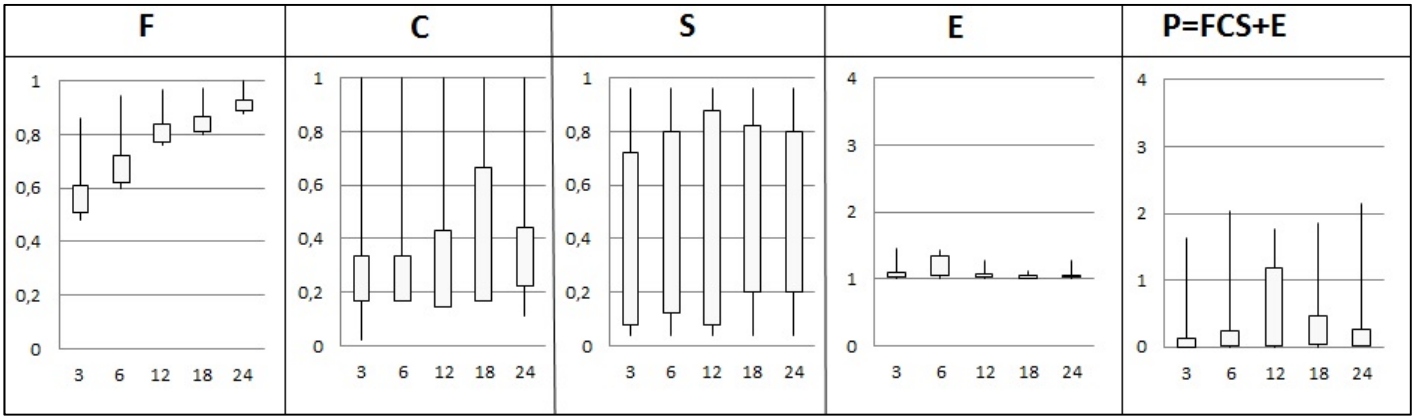


Figure 44 : Variations des mesures d'intérêt F, C, S, E et de leur combinaison P calculées sur l'ensemble des motifs obtenus pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec les grilles DCE (les extrémités du rectangle représentent les 1ers et 3èmes quartiles, les extrémités des lignes les minima et maxima)

3.2 Caractérisation des altérations de l'ensemble des motifs obtenus par longueurs de séquences

Nous avons fait le choix de limiter les altérations les plus abondantes dans l'objectif de voir apparaitre d'autres altérations dans les motifs. Malgré cela, les motifs extraits avec les grilles DCE sont très peu variés. Leurs items correspondent aux altérations du Bilan de l'Oxygène (BILO2) en très bon, bon ou état moyen, des polluants spécifiques (POSPE) en bon état, uniquement pour 3, 6, 12 et 18 mois, puis les POSPE uniquement en mauvais état à 18 mois, et les substances prioritaires et dangereuses prioritaires (SDP) en mauvais état. Nous utiliserons l'appellation substances prioritaires pour désigner les substances prioritaires et dangereuses prioritaires dans la suite du chapitre. Enfin, à 3 mois et 6 mois, quelques motifs ont des items nutriments en bon état (NUTRI_bleu). Les pourcentages d'apparitions dans les données d'entrée et dans les motifs sont donnés dans le Tableau 10 et leur nombre par la Figure 45. Les items SDP sont les plus nombreux, jusqu'à 75% pour 24 mois. Le pourcentage des items BILO2 varie entre 23,3% à 3 mois et 40,9% à 6 mois. Celui des items POSPE représente 21,9% des items totaux à 3 mois, puis est proche de 6% à 6 et 12 mois, et inférieur à 1% à 18 et 24 mois.

Tableau 10 : Pourcentages d'altérations présents dans les données d'entrée et dans l'ensemble des motifs obtenus pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec les grilles DCE

Altérations	Données d'entrée	Motifs				
		3 mois	6 mois	12 mois	18 mois	24 mois
TEMP	19,4	0,0	0,0	0,0	0,0	0,0
ACID	19,3	0,0	0,0	0,0	0,0	0,0
BILO2	19,5	23,3	40,9	27,5	30,9	24,6
NUTRI	17,3	0,8	6,0	0,2	0,0	0,0
POSPE	11,7	21,9	6,5	6,2	0,1	0,3
SDP	12,8	54,0	46,6	66,1	69,0	75,0

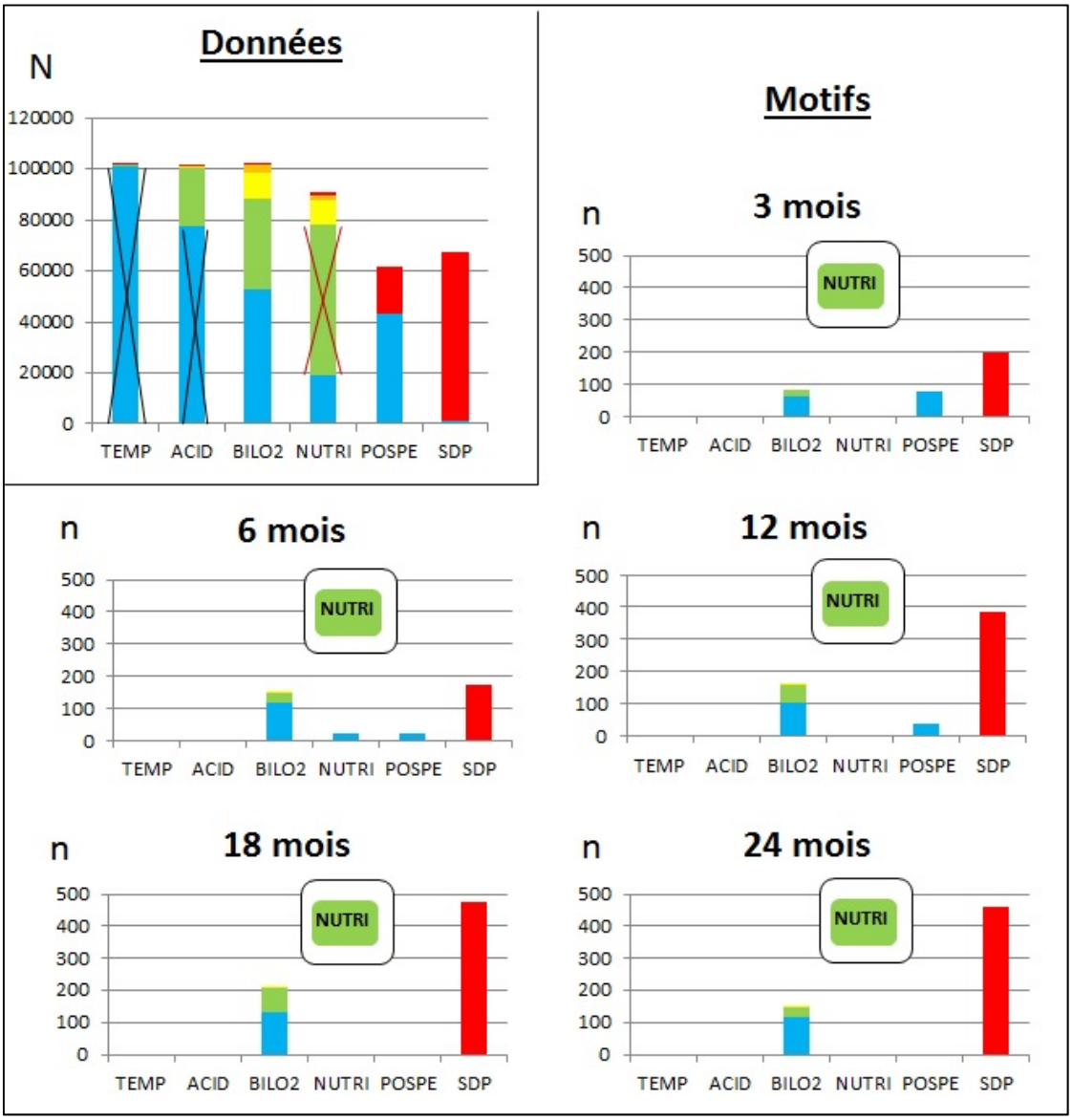


Figure 45 : Nombre d'altérations, par classe de qualité, présentes dans les données d'entrée et dans l'ensemble des motifs obtenus pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec les grilles DCE (les altérations d'entrée barrées sont celles retirées avant extraction des motifs car trop abondantes dans les données d'entrée – en noir – ou trop abondantes dans les premières extractions de motifs – en rouge)

3.3 Motifs caractéristiques obtenus par longueurs de séquences

Les cinq motifs sélectionnés par contexte sur la base de leur combinaison P la plus élevée, ainsi que leurs caractéristiques (mesures d'intérêt, items) sont disponibles en annexes 5a à 5e par longueurs de séquences de 3 à 24 mois.

Parmi ces cinq motifs sélectionnés, nous n'avons retenu ci-après que les motifs dominants par indice : il s'agit de ceux extraits pour l'IBD en très bon et bon états (Figure 46 et Figure 47), pour l'IPR en très bon état (Figure 48), pour l'IBMR en états très bon, médiocre et mauvais (Figure 49, Figure 50 et Figure 51), et pour l'IBGN en mauvais état (Figure 52). Aucun motif dominant n'a été trouvé pour les contextes de l'I2M2.

Pour le contexte IBD bleu (Figure 46) sont observés des motifs dominants pour chaque longueur de séquences. Les items sont essentiellement les nutriments en très bon état (NUTRI_Bleu) associés aux items du bilan oxygène en très bon état (BILO2_Bleu) pour 3 et 6 mois. A 12 mois, ils sont présents mais dans des motifs séparés. A 18 et 24 mois, seul demeure l'item BILO2_Bleu qui se répète 3 ou 4 fois respectivement pour 24 et 18 mois. Le motif le plus complexe est le n° 54 obtenu à 6 mois avec 4 itemssets et 5 items.

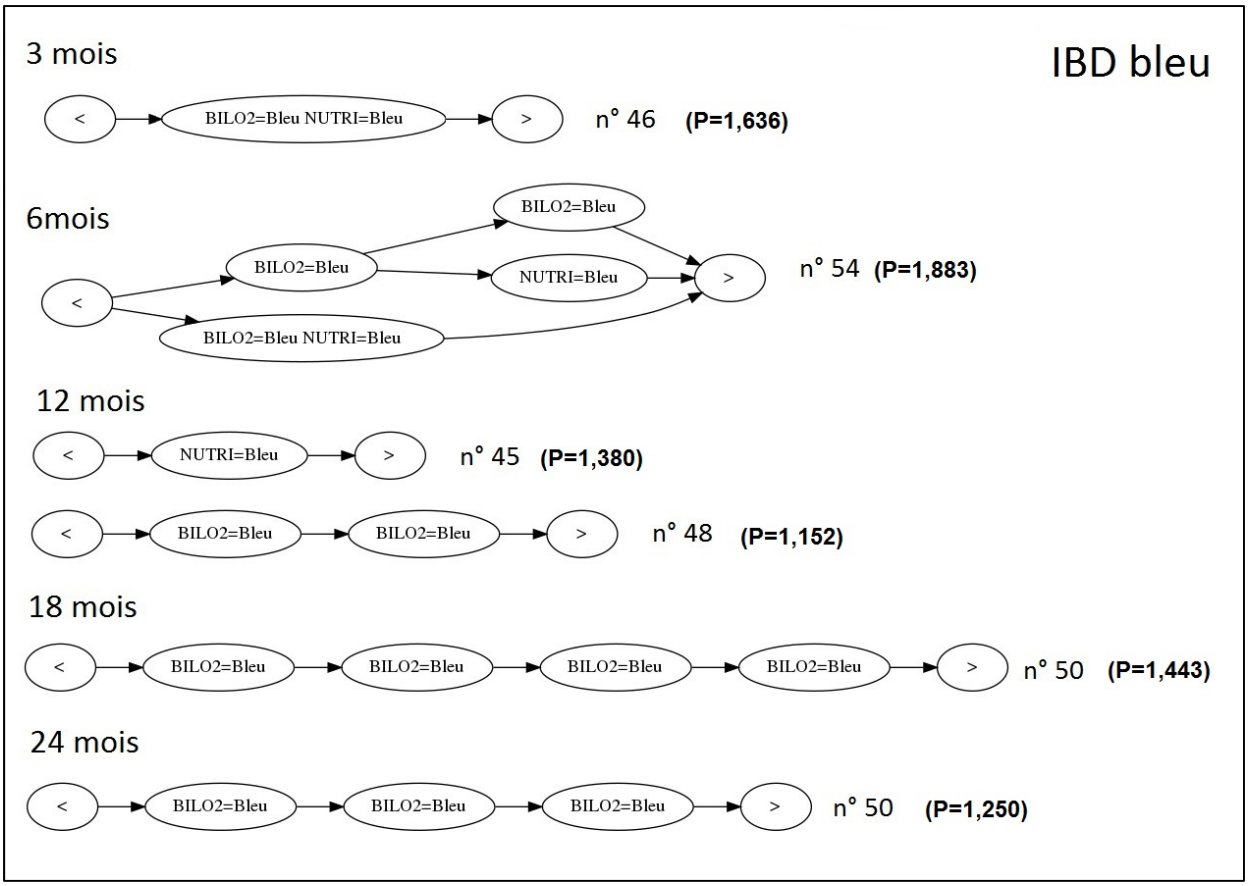


Figure 46 : Motifs extraits pour le contexte IBD en très bon état, choisis parmi les 5 motifs sélectionnés sur la base de leur combinaison P (indiquée entre parenthèses) la plus élevée, pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec les grilles DCE

Le contexte de l'IBD en bon état n'avait pas de motif dominant pour les extractions faites avec le SEQ-eau, alors qu'avec les grilles DCE il a des motifs dominants pour toutes les longueurs de séquences excepté 12 mois (Figure 47). Le seul item en commun avec les motifs extraits pour le contexte de l'IBD en très bon état est le bilan en oxygène en très bon état (BILO2_Bleu) qui est présent dans les motifs dominants pour toutes les longueurs de séquences, mais en association avec un à trois items de substances prioritaires en mauvais état (SDP_Rouge) dans tous les motifs retenus pour les longueurs de séquences 3 et 6 mois, et avec les polluants spécifiques en très bon état (POSPE_Bleu) dans les motifs n° 93 à 3 mois et 89 à 6 mois. L'Item BILO2_Bleu est associé à un item de cette même altération en bon état (BILO2_Vert) pour les motifs sélectionnés à 18 et 24 mois (n° 78 et 80).

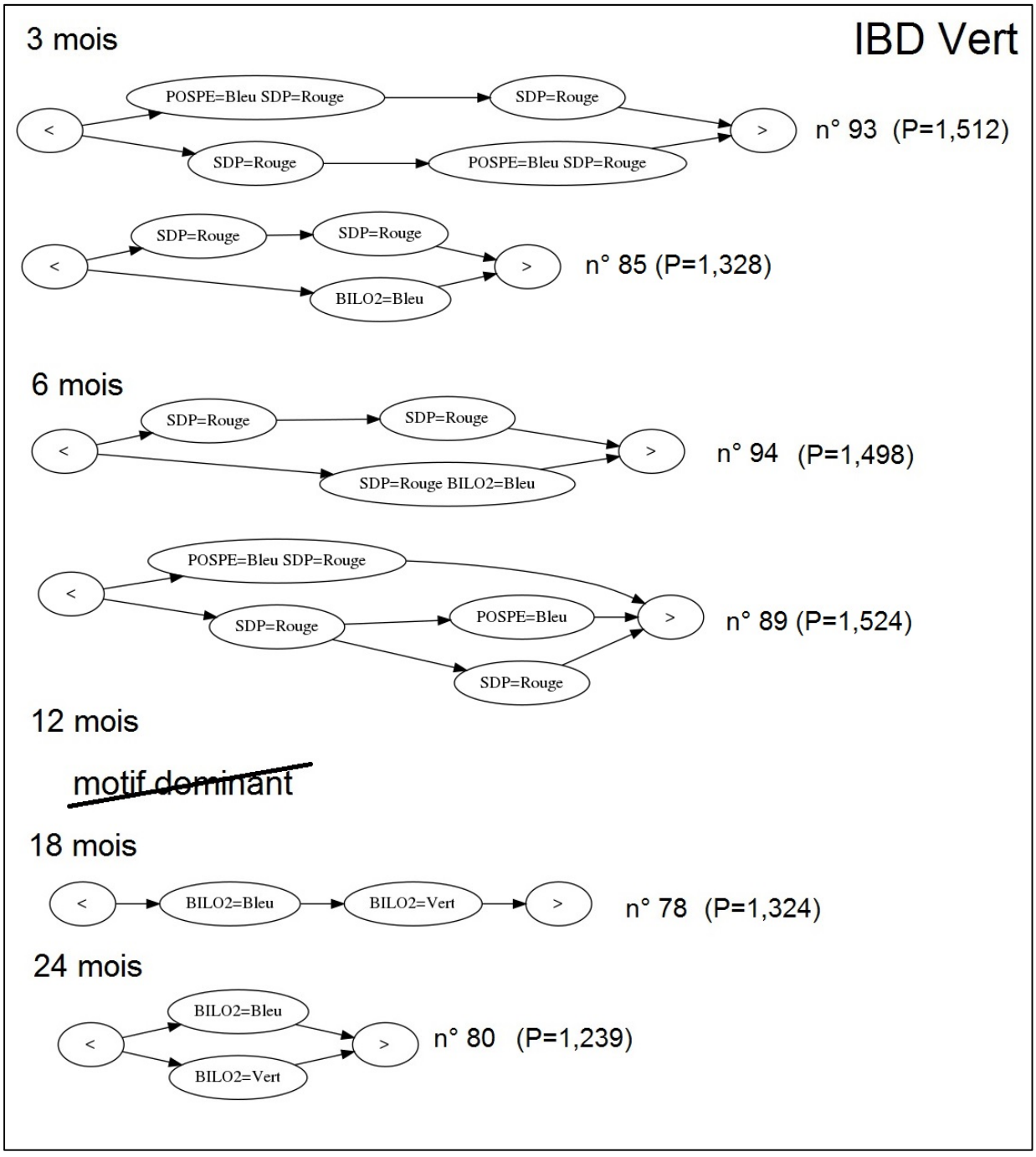


Figure 47 : Motifs extraits pour le contexte IBD en bon état, choisis parmi les 5 motifs sélectionnés sur la base de leur combinaison P (indiquée entre parenthèses) la plus élevée, pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec les grilles DCE

Les motifs dominants extraits pour le contexte de l'IPR en très bon état sont composés principalement de l'item du bilan de l'oxygène en très bon état (BILO2_Bleu) répétés 2 ou 3 fois à 3 et 6 mois (Figure 48). Y sont également associés les items des polluants spécifiques en très bon état (POSPE_Bleu) et des substances prioritaires en mauvais état (SDP_Rouge). Il n'y a plus de motif dominant dans les extractions faites à partir des longueurs de séquences de 12 mois.

Pour les deux contextes de l'IBD en très bon et bon état, les motifs obtenus pour les longueurs de séquences 18 et 24 mois sont limités à quatre ou deux items, et toujours moins complexes que ceux obtenus à 6 mois, voire à 3 mois dans le cas de l'IBD_Bleu. Le minimum de fréquence supérieur à 0,60 à partir de 12 mois et le nombre important de stations disponibles pour ces deux contextes – respectivement 2 297 et 2 719 – doivent en être responsables. Par contre, lorsque le nombre de stations est plus réduit pour les contextes suivants de l'IBMR en états très bon, médiocre et mauvais et l'IBGN en mauvais état – respectivement 347, 116, 19 et 23 stations –, les motifs obtenus malgré la même limite minimale de fréquence sont plus complexes : les motifs concernés ont de 5 à 9 items (motif n° 171 pour IMBR_Rouge, Figure 51). Mais la combinaison d'une fréquence minimale élevée et d'un grand nombre de stations disponibles ne correspond pas au cas de l'IPR_Bleu, pour lequel il n'y a pas de de motif dominant dans les extractions faites à partir de 12 mois, alors que le nombre de stations disponibles pour ce contexte est de 274, ce qui est proche des valeurs pour les contextes IBMR_Orange et IBMR_Bleu.

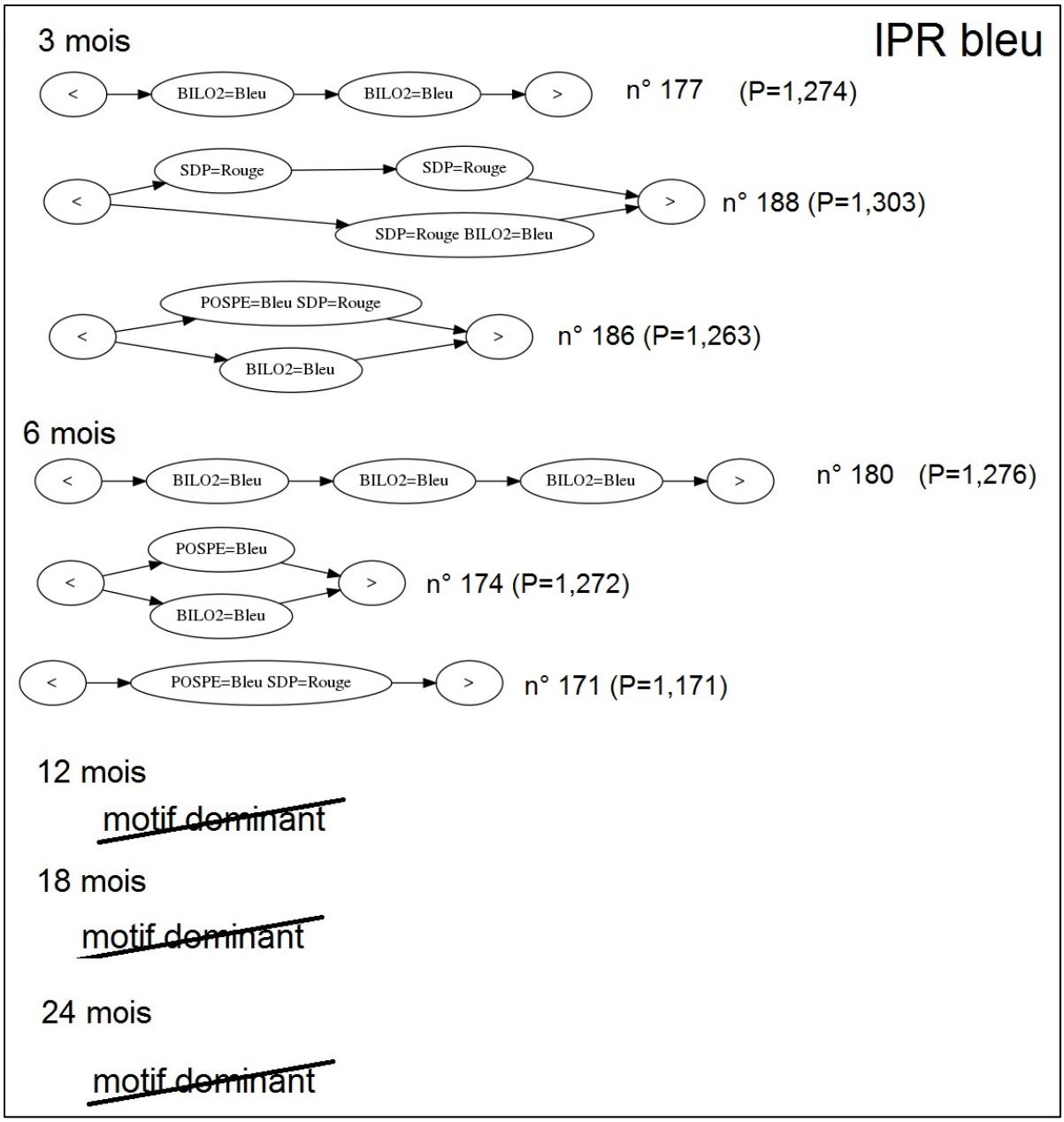


Figure 48 : Motifs extraits pour le contexte IPR en très bon état, choisis parmi les 5 motifs sélectionnés sur la base de leur combinaison P (indiquée entre parenthèses) la plus élevée, pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec les grilles DCE

Pour les longueurs de séquences de 12, 18 et 24 mois, les motifs sélectionnés pour le contexte de l'IBMR en très bon état sont tous dominants et contiennent tous les deux les mêmes items : le bilan de l'oxygène en très bon état (BILO2_Bleu) et les substances prioritaires en mauvais état (SDP_Rouge) (Figure 49). Ils ont au maximum 6 items, et ce sont les SDP_Rouge les plus nombreux, de 4 à 5 dans les plus grands motifs. Il n'y pas de motif dominant extrait pour 3 mois. A 6 mois, l'item

BILO2_Bleu est associé avec un item de la même altération mais de bon état (BILO2_Vert).

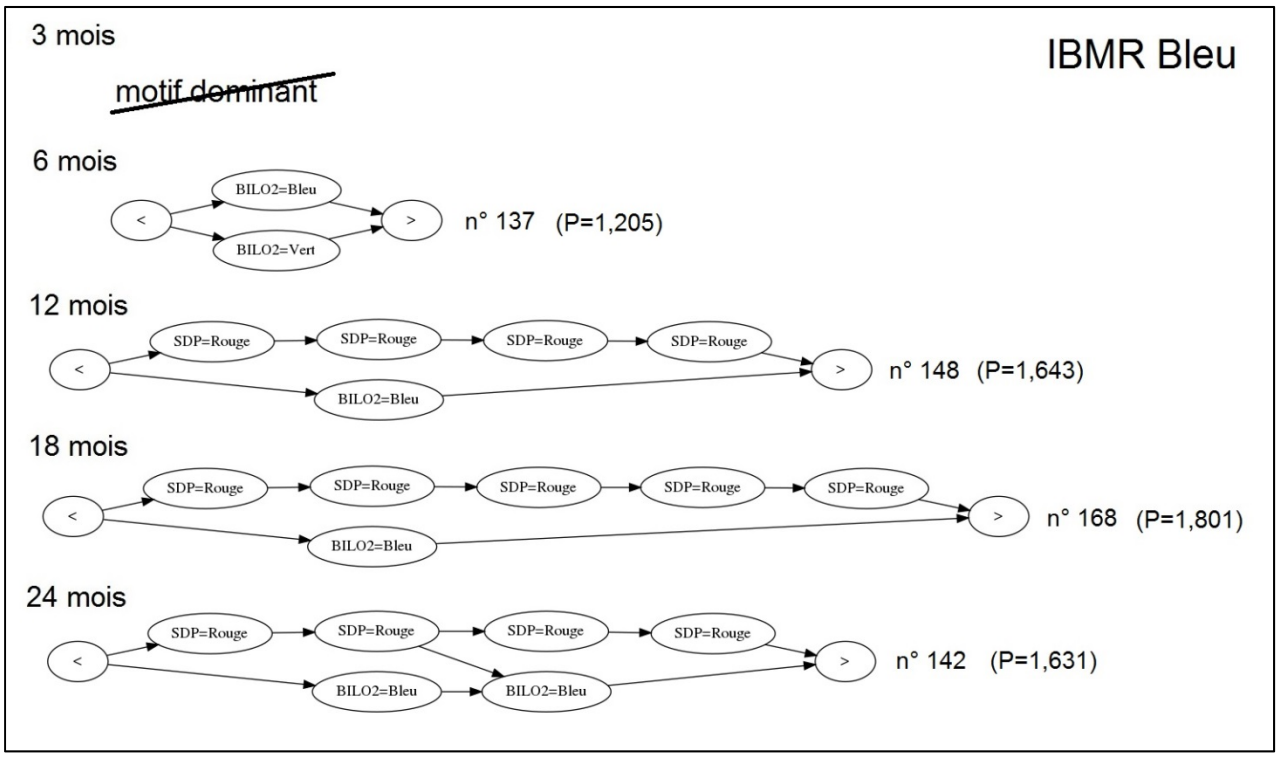


Figure 49 : Motifs extraits pour le contexte IBMR en très bon état, choisis parmi les 5 motifs sélectionnés sur la base de leur combinaison P (indiquée entre parenthèses) la plus élevée, pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec les grilles DCE

Les motifs dominants extraits pour le contexte de l'IBMR d'état médiocre (Figure 50) et mauvais (Figure 51) sont composés tous deux des items du bilan oxygène en bon état (BILO2_Vert) et des substances prioritaires en mauvais état (SDP_Rouge). Par contre, pour le contexte IBMR_Orange l'item des polluants spécifiques est présent en très bon état (POSPE_Bleu) dans quelques motifs extraits pour 12 mois (dont le motif n° 200, Figure 50) alors qu'il est remplacé par l'item de cette altération en mauvais état (POSPE_Rouge) pour le contexte de l'IBMR_Rouge (motif n° 171, Figure 51). L'item bilan oxygène en état moyen (BILO2_Jaune) apparait ponctuellement dans le motif n° 203 (Figure 51) extrait pour l'IBMR_Rouge à 18 mois.

Il y a peu de différences entre les motifs dominants sélectionnés pour le contexte de l'IBMR en mauvais état et ceux pour le contexte de l'IBGN en mauvais état (Figure 52), mis à part que l'item des polluants spécifiques en mauvais état (POSPE_Rouge) n'apparait pas. Par contre, il est à noter que trois items des substances prioritaires (SDP_Rouge) se succédant sont présents dès 3 mois. Il y en a six à 6 mois ainsi qu'à 18 mois. Cet item doit englober les altérations du SEQ-eau des micropolluants minéraux (MPMI) et organiques hors pesticides (MPOR) qui étaient bien représentés dans les motifs extraits pour ce contexte avec le SEQ-eau, quelles que soient les longueurs de séquences.

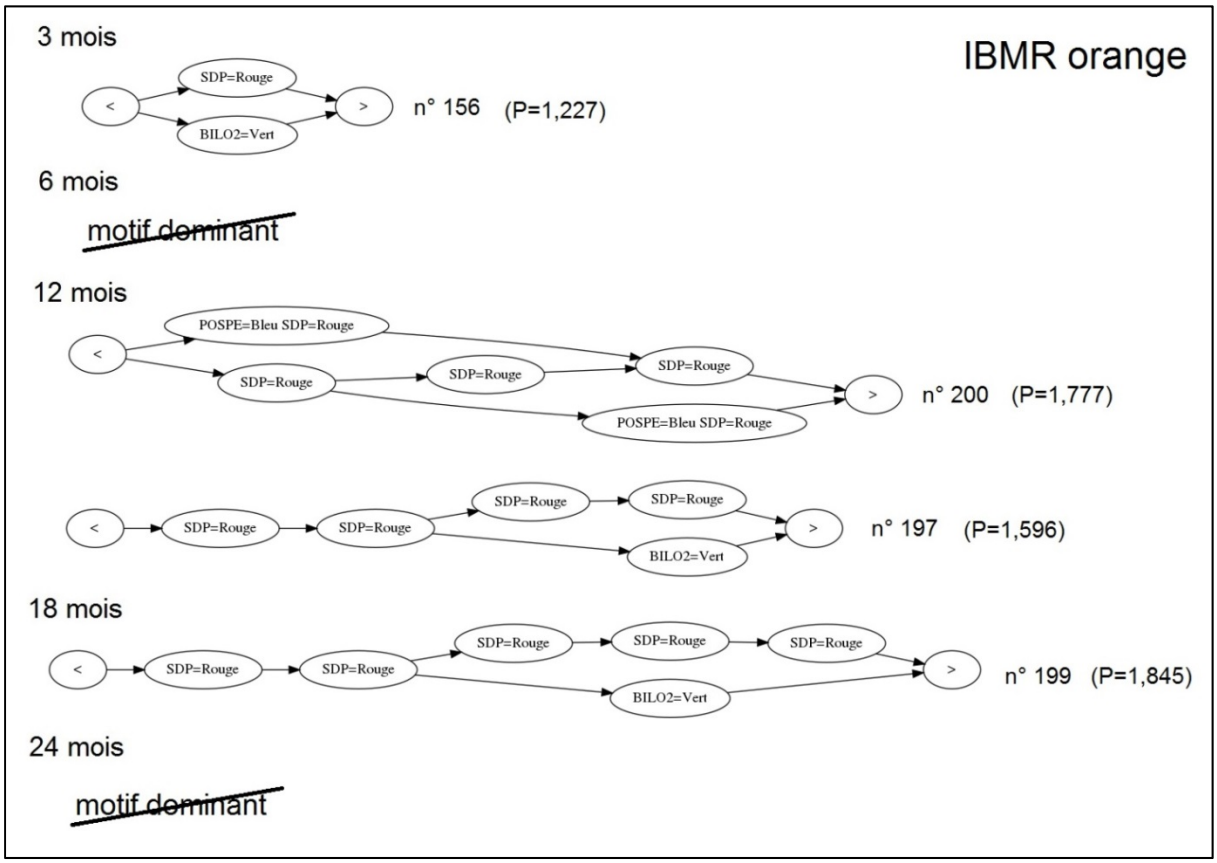


Figure 50 : Motifs extraits pour le contexte IBMR en état médiocre, choisis parmi les 5 motifs sélectionnés sur la base de leur combinaison P (indiquée entre parenthèses) la plus élevée, pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec les grilles DCE

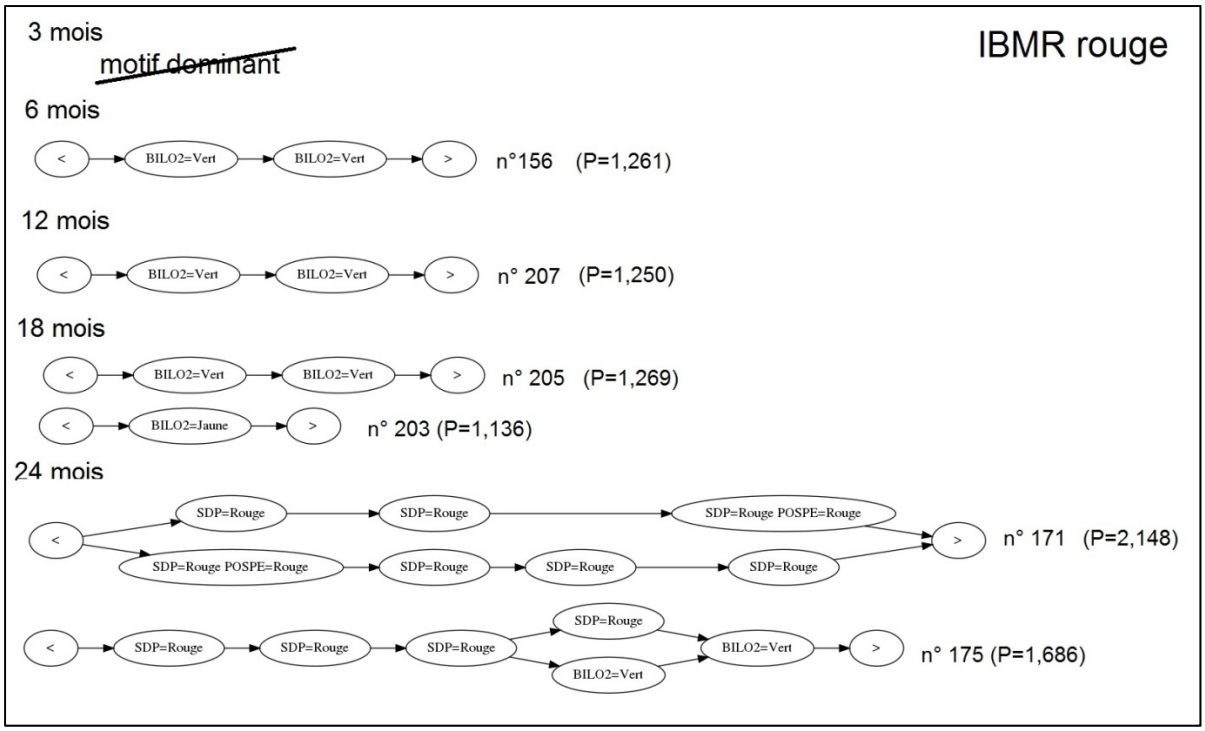


Figure 51 : Motifs extraits pour le contexte IBMR en mauvais état, choisis parmi les 5 motifs sélectionnés sur la base de leur combinaison P (indiquée entre parenthèses) la plus élevée, pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec les grilles DCE

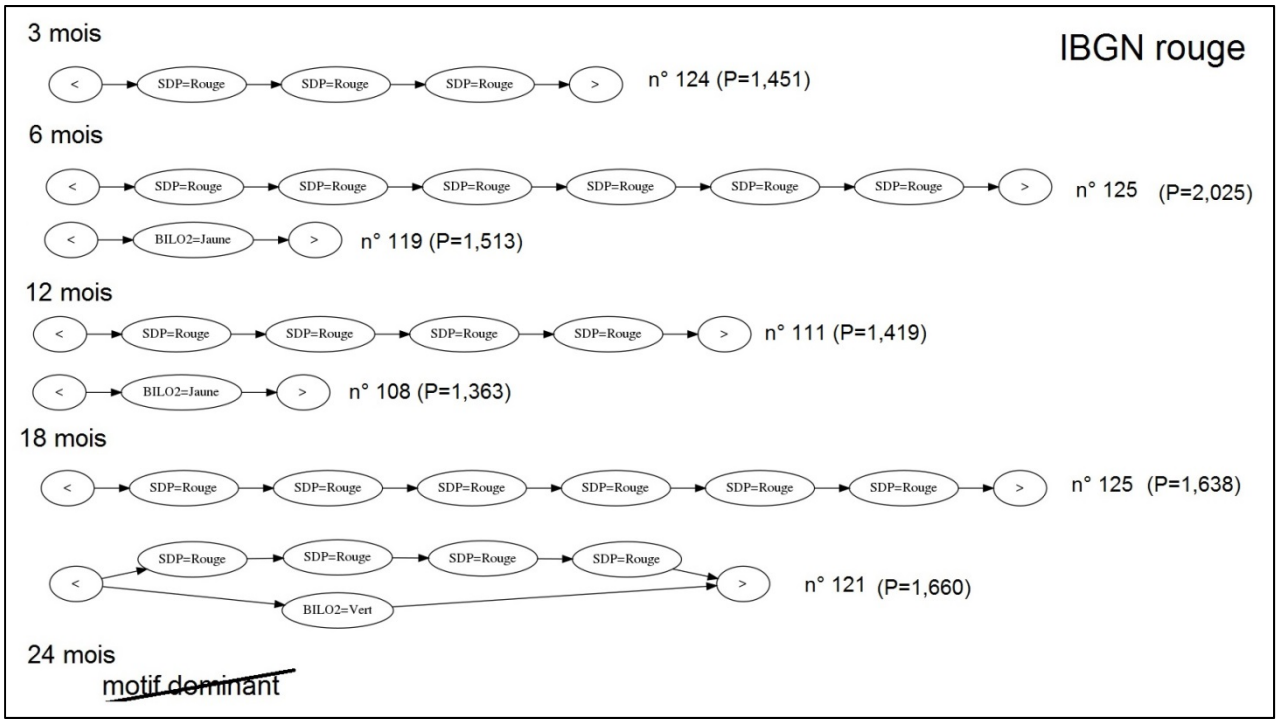


Figure 52 : Motifs extraits pour le contexte IBGN en mauvais état, choisis parmi les 5 motifs sélectionnés sur la base de leur combinaison P (indiquée entre parenthèses) la plus élevée, pour les longueurs de séquences 3, 6, 12, 18 et 24 mois, avec les grilles DCE

4 Discussion

La question posée dans ce chapitre était : **les motifs extraits pour différentes longueurs de séquences sont-ils différents en fonction des indices biologiques?**

Parmi les altérations que nous utilisons, que ce soit avec le SEQ-eau ou les grilles DCE, nous distinguons ci-après celles qui sont dans les classes très bonne et bonne, et celles désignant des altérations au sens de dégradations qui sont dans les classes moyenne, médiocre et mauvaise.

Nous avons montré dans le chapitre précédent que les indices biologiques en très bon état étaient précédés de motifs composés majoritairement d'altérations dans les classes très bonne et bonne. Pour les altérations du SEQ-eau, c'est ce que nous observons principalement pour les phosphates (PHOS_bleu) pour tous les indices biologiques en très bon état et spécialement pour l'IBD. Pour ce contexte, dans les motifs sélectionnés, le nombre d'items PHOS_Bleu augmente progressivement de 1 à 7 successions lorsque l'on passe de 3 à 24 mois. C'est aussi le cas, pour l'item de l'altération acidité en bon état (ACID_Vert) uniquement pour l'indice IBD en très bon état. Cet item est présent en nombre important dans l'ensemble des motifs extraits de 3 à 24 mois. Dans les motifs sélectionnés pour ce contexte, cet item est en association avec les phosphates en très bon état pour toutes les longueurs de séquences. C'est aussi le cas, dans une moindre mesure, pour les altérations matières organiques et oxydables et nitrates en très bon état (MOOX_bleu et NITR_Bleu) : leurs items sont présents dans les motifs pour tous les indices biologiques en très bon état dès 3 mois. Leur nombre augmente progressivement avec l'augmentation des séquences. Nous utiliserons l'appellation matières organiques pour désigner les matières organiques et oxydables dans la suite du chapitre. L'item MOOX_Bleu se répète deux fois à 3 mois, puis trois fois à partir de 6 mois, dans les motifs sélectionnés pour le contexte IPR en très bon état. L'item NITR_Bleu est associé avec l'item PHOS_Bleu dans les motifs sélectionnés pour l'IBD en très bon état pour la longueur de séquences 3 mois. Pour les altérations de la grille DCE, ce sont les altérations du bilan en oxygène et des polluants spécifiques

en très bon état qui sont récurrents dans les motifs sélectionnés pour les indices IBMR et IPR en très bon état.

D'une manière générale, quel que soit l'indice biologique, le nombre d'apparitions des items aura tendance à augmenter avec l'allongement des séquences et diminuer lorsque la fréquence minimale augmente. Or, pour gérer les limites de notre algorithme PRESTOR, nous avons augmenté cette fréquence minimale en même temps que l'augmentation des longueurs de séquences. Pour les altérations en état moyen et au-delà, nous faisons l'hypothèse de profils d'apparition d'items assez semblables d'une longueur de séquences à l'autre, pour des altérations qui auraient un effet dès 3 mois sur le groupe biologique considéré ; et à l'inverse pour des altérations qui auraient un effet retardé, des profils différenciés avec des apparitions d'abord absentes ou faibles pour des longueurs de séquences courtes puis plus importantes pour les longueurs de séquences plus longues. Nous faisons également l'hypothèse que les items des altérations ayant un impact immédiat sont observées en nombre dans les motifs extraits dès 3 mois ; et à l'inverse que les items des altérations ayant un effet retardé n'apparaissent que dans les motifs extraits pour des longueurs de séquences plus longues.

Berenzen et al. (2001) ont également montré que l'ammonium pouvait avoir un effet toxique à partir 3 mg/L pour le crustacé *Gammarus pulex* (résultats obtenus en microcosmes), à partir de 30 mg/L pour le mollusque *Radix ovata* et le Trichoptère *Limnephilus lunatus*. La concentration de 3 mg/L d'ammonium correspond à une qualité médiocre du SEQ-eau (seuils de la classe médiocre de NH_4^+: [2 ;5 mg/L[) et entraine l'apparition de 0,91 mg/L de nitrites, ce qui correspond aussi à une qualité médiocre (seuils SEQ-eau de NO_2^- :[0,5 ; 1 mg/L[), et donc par conséquent induit une qualité médiocre pour l'altération matières azotés hors nitrates (AZOT) que composent ces deux paramètres. Nous n'avons pas observé cette altération dans cette qualité dans les motifs extraits. Par contre, l'item de cette altération de qualité moyenne (AZOT_Jaune) est comptabilisé dans les motifs extraits pour le contexte de l'I2M2 en mauvais état, et ce quelles que soient les longueurs de séquences.

Mais ce sont surtout les effets toxiques des micropolluants qui sont connus et peuvent être immédiats sur les différents groupes biologiques aquatiques. Dans le cas de contaminations toxiques aigües, les mortalités des espèces aquatiques

peuvent être très rapides. Une extinction massive des poissons du Rhin mais aussi des macro-invertébrés (Van Urk et al., 1993) a eu lieu en quelques heures après l'incendie de l'usine Sandoz de Bâle, en 1986, qui a pollué le fleuve principalement en insecticides organochlorés et organophosphorés (1 200 tonnes) et mercure (2 tonnes) (Gautier, 2001). Les doses létales pour 50% d'une population (DL50) des principaux micropolluants organiques et métaux lourds sont connus pour les principaux groupes de macro-invertébrés (Wogram et Liess, 2001). Ces valeurs sont très variables et les effets cocktails inconnus. Les effets toxiques sur les macrophytes sont moins étudiés, même s'il existe des tests en laboratoire sur les lentilles d'eau. La capacité de certaines espèces à stocker, voire métaboliser certaines molécules les rend plus résistances, mais des dysfonctionnements physiologiques tels qu'un retard de croissance des racines ont été observés chez des macrophytes suite des expositions à des pesticides (Arts et al., 2008). Pour tous les indices biologiques, nous observons dans les motifs extraits dès 3 mois, l'apparition des items de l'altération pesticides de qualité médiocre et mauvaise (PEST_Orange et PEST_Rouge) du SEQ-eau ou de l'altération des substances prioritaires en mauvais état (SDP_Rouge) des grilles DCE. Il ressort donc des analyses que les micropolluants ont une incidence forte sur le degré d'altération révélé par les indices. Par contre, ces items sont peu discriminants vis-à-vis des états biologiques, mis à part pour l'IBMR et l'IPR en mauvais état pour la longueur de séquences 24 mois. Les items de l'altération des PCB de qualité moyenne (PCB_Jaune) ne sont comptabilisés que pour les extractions faites à 3 mois pour les indices I2M2, IBGN et IPR, mais pour ce dernier, ils ne sont pas discriminants au regard de ses états biologiques. Leur absence dans les extractions faites pour les autres longueurs de séquences, ayant des fréquences minimales supérieures à 0,30, s'explique par le faible taux de complétude de cette altération dans les données d'entrée (30% contre en moyenne 50% pour les autres micropolluants), comme montré dans le chapitre précédent (Chapitre III, 2.6). L'I2M2 en mauvais état est le seul contexte pour lequel cet item apparait dans un des motifs sélectionnés. C'est seulement pour l'IBGN en mauvais état que sont comptabilisés, dès 3 mois et jusqu'à 24 mois, les items des micropolluants organiques hors pesticides de qualité moyenne (MPOR_Jaune) et des micropolluants minéraux de qualité médicore (MPMI_Orange). Ces items sont aussi présents dans tous les motifs sélectionnés pour ce contexte, quelles que soient les longueurs de séquences. La récurrence des

altérations des micropolluants dans les motifs semble indiquer que leur effet est fort et leur faible pouvoir discriminant qu'ils sont très largement répandus dans les hydrosystèmes.

En deçà des doses létales, les temps de réponses des organismes dépendent non seulement du groupe biologique mais aussi des types de polluants, de leur concentration et de la présence d'autres pressions, soit par effet additif, antagoniste, synergique entre polluants, soit par effets cumulés entre différentes sources de perturbation. Liess et Beketov (2011) ont montré, en mésocosmes, que l'abondance totale et la richesse d'un peuplement de macro-invertébrés aquatiques étaient significativement altérées 3 semaines après exposition à des doses de 100µg/L de l'insecticide thiaclopride. Pringle (1990) a mis en évidence, en condition expérimentale, des réponses significatives de diatomées benthiques (multiplication du biovolume) en 20 jours lors d'enrichissement du milieu en nutriments (de 13 à 25 mg/l de NO_3^- et de 0,25 à 1, 2 mg/L de PO_4^{3-}). Plusieurs auteurs ont montré que la pollution par les micropolluants, et les pesticides en particulier, était un problème mondial et menaçait la biodiversité aquatique (Altenburger et al., 2015; Malaj et al., 2014; Stehle and Schulz, 2015). D'après Malaj et al. (2014), 42 % des sites surveillés en Europe sont exposés à des pollutions chroniques ayant des effets sur les espèces aquatiques (étude réalisée sur 8 200 sites).

Enfin, le temps de recolonisation d'un écosystème dulcicole après une perturbation majeure dépend des cycles de vie des espèces. Chez les diatomées, le temps entre deux générations est de l'ordre de quelques dizaines d'heures à quelques jours, mais varient en fonction des conditions environnementales (Elbrächter, 1977). Chez les macro-invertébrés, le temps de recolonisation est de quelques semaines pour les espèces polyvoltines, ce qui plus rapide pour les espèces monovoltines. Le critère de la fréquence relative des organismes polyvoltins au sein des habitats dominants est d'ailleurs une des métriques retenues dans l'I2M2 (Mondy et al., 2012). Van Urk et al., (1993) ont montré, sur quelques espèces de chironomidés et un trichoptère (*Hydropsyche contubernalis*), que deux générations – soit quelques semaines pour les chironomidés – étaient nécessaires pour que des espèces disparues suite à une pollution accidentelle massive en insecticides (incendie de Sandoz, à Bâle, 1986) se réinstallent durablement, sous réserve que l'ensemble des conditions environnementales soient retrouvées. En conditions

naturelles, Schulz et Liess (1999) ont montré qu'il fallait entre 3 à 6 mois à plusieurs espèces d'invertébrés (planaires, gastéropodes, crustacés, trichoptère, diptères) pour recoloniser un écosystème pollué aux pesticides. Winemiller et Rose (1992) répartissent les poissons en trois groupes en fonction de leur maturité sexuelle : ceux dont la maturité est inférieure ou égale à un an, ceux dont la maturité est supérieure ou égale à un an, sous-répartis entre ceux qui prennent soin de leur jeunes et les autres. La recolonisation animale après une perturbation dépend de la rémanence de celle-ci mais également de la proximité de populations source susceptibles de rejoindre le milieu. Les macrophytes ont plusieurs stratégies annuelles de dispersion par reproduction sexuelle et/ou végétative, de plus, la banque de graines des sédiments assure leur possible résilience lorsque les conditions environnementales y sont favorables (Combroux et al., 2001). Dans le cadre des restaurations dynamiques d'anciens bras déconnectés du Rhin, Meyer et al. (2013) ont montré qu'à conditions physico-chimiques équivalentes, une année suffisait à l'installation d'un peuplement de macrophytes dans de bonnes conditions hydro-géomorphologiques. L'extraction des motifs pour l'IBMR serait à recommander à partir de longueur de séquences de 12 mois.

Pour les altérations des macro-polluants dans les classes moyenne à mauvaise, les motifs que nous avons extraits pour l'IBMR et l'IBGN se distinguent de ceux obtenus pour l'IBD en fonction des longueurs de séquences. En effet, dans l'ensemble des motifs de l'IBMR et de l'IBGN, les items d'une partie des altérations sont absents ou en très faible nombre pour les extractions aux longueurs de séquences courtes, 3 et 6 mois, et importantes pour les extractions aux longueurs de séquences plus longues, au-delà de 12 mois. Avec les altérations du SEQ-eau, c'est le cas pour les matières organiques (MOOX) et les nitrates (NITR) de qualité moyenne, pour les états médiocre et mauvais de l'IBMR et l'état mauvais de l'IBGN. Alors que ces items sont présents dans les motifs de l'IBD d'état moyen et au-delà dès les extractions à 3 mois. C'est aussi le cas pour les particules en suspension (PAES), qui apparaissent pour l'IBMR en mauvais état surtout à partir des longueurs de séquences de 18 mois. A l'inverse les nombres d'items des altérations matières azotées hors nitrates (AZOT) et matières phosphorées (PHOS) de qualité moyenne sont assez constants quelle que soit la longueur de séquences, dès 3 mois, pour l'IBMR en états médiocre et mauvais. Robach et al. (1996) ont montré que les

macrophytes répondaient rapidement aux taux de nutriments : phosphates et plus à l'ammonium qu'aux nitrates. D'après leurs travaux, la communauté de macrophytes caractéristique des milieux les plus eutrophes, en plaine d'Alsace, étaient dans des eaux dont la concentration moyenne en ammoniums était de 0,33 mg/L, ce qui correspond à une bonne qualité pour ce paramètre que ce soit avec le SEQ-eau ou les grilles DCE (]0,1 ; 0,5]). La concentration moyenne en phosphates était de 0,59 mg/L ce qui correspond à une qualité moyenne pour ce paramètre que ce soit avec le SEQ-eau ou les grilles DCE (]0,5 ; 1]). Les motifs sélectionnés et extraits pour 12 mois et plus pour les contextes de l'IBGN et l'IBMR mauvais, sont composés d'au moins trois items de deux macro-polluants dégradés, ou au moins un item de macropolluant associé à plusieurs items de micropolluants dégradés. Pour les altérations des grilles DCE, les motifs sélectionnés sont composés de l'item du bilan en oxygène en bon état (BILO2_Vert) associé à plusieurs items de micropolluants en mauvais état à partir de 18 mois pour l'IBGN et de 24 mois pour l'IBMR, en mauvais état.

Ainsi nous avons observé des différences dans les motifs extraits à différentes longueurs de séquences pour les indices IBD, IBGN et IBMR. Il est probable que la durée de vie des organismes utilisés pour ces indices ait une influence sur la relation qui peut être établie entre un état biologique et la chronique des événements physico-chimiques qui l'ont précédé.

Pour les indices IPR et I2M2, nous n'avons pas observé de différences significatives des motifs extraits aux différentes longueurs de séquences. Ces motifs étaient tous assez courts avec des altérations similaires. Pour les poissons, extraire des motifs limités aux seules altérations physico-chimiques ne semble pas suffisant et il faudrait y intégrer les pressions hyromorphologiques auxquelles ils sont reconnus comme sensibles (Reyjol et al., 2008; Villeneuve et al., 2015). Pour l'I2M2, les seuils de discrétisation datant de 2012 que nous utilisons peuvent expliquer le peu de motifs suffisamment complexes et pertinents que nous avons trouvés pour cet indice, contrairement aux motifs pour l'IBGN. Depuis 2012, les seuils de l'I2M2 ont été affinés et il en existe à présent des nouveaux qui seraient à prendre en compte.

CHAPITRE V : Comment sélectionner des motifs caractéristiques à l'échelle d'une hydro-écorégion ?

Nous avons développé le programme, PRESTOR sous licence libre, qui permet à un opérateur d'extraire des motifs pour un indice biologique dans un état donné, le contexte. L'opérateur peut choisir différents critères d'extraction, dont le territoire – la France entière ou une Hydro-Eco-Région (HER) –, la longueur des séquences (stations-dates), c'est-à-dire l'intervalle de temps précédant l'état biologique (Chapitre III).

Les algorithmes de fouille de données fournissent en général un grand nombre de résultats aussi il est indispensable de leur associer des mesures d'intérêt, permettant de sélectionner les résultats les plus pertinents. Dans ce but, au chapitre III, nous avons proposé quatre mesures d'intérêts à associer à chaque motif: la fréquence, la complexité, la singularité et l'émergence, ainsi que leur combinaison. Rappelons que la **fréquence** d'un motif est le rapport entre le nombre de séquences vérifiant ce motif et l'ensemble des séquences ayant le contexte choisi. Le nombre de séquences vérifiant le motif est le support du motif. La **complexité** d'un motif correspond au nombre d'altérations qu'il contient comparé au nombre maximal d'altérations trouvées dans l'ensemble des motifs de l'extraction. La **singularité** correspond au degré de spécificité du motif pour son contexte d'extraction, un motif identique pouvant être extrait pour différents contextes. L'**émergence** n'est calculée que pour les motifs dominants dans leur contexte d'extraction, c'est-à-dire avec la fréquence maximale. Elle correspond au rapport entre cette fréquence et la deuxième fréquence la plus importante pour ce même motif dans un autre contexte. Nous cherchons, dans ce chapitre, à affiner, automatiser la sélection de motifs caractéristiques parmi les motifs extraits et à compléter la sélection des motifs à l'échelle d'une hydro-écorégion (HER) telles de définies par Wasson et al. (2004).

Par ailleurs, à l'échelle nationale, PRESTOR a permis d'extraire des motifs pour des longueurs de séquences maximales de 24 mois.

Ce chapitre a pour objectif de répondre à trois questions, les deux premières étant liées à la méthode développée, la dernière étant une des questions majeures des gestionnaires de rivières dont l'état écologique devrait être restauré :

1. **réduire l'échelle spatiale, de nationale à celle d'une HER, permet-elle de d'allonger l'intervalle de temps d'extraction de motifs ?**
2. **comment affiner et automatiser la sélection des motifs les plus pertinents ?**
3. **existe-t-il des successions d'altérations physico-chimiques caractéristiques d'un changement d'état biologique à l'échelle d'une HER?**

1 Matériel et méthodes

Nous avons retenu l'HER 18, qui correspond à la plaine d'Alsace, dont le nombre de stations est limité à quelques dizaines et que nous connaissons bien. Nous utilisons les données françaises nationales acquises sur le réseau de contrôle et de surveillance de cette HER, de 2007 à 2013, soit 72 mois pour les indices biologiques et de 2007 à 2012, soit 60 mois pour les paramètres physico-chimiques. Nous nous limitons aux seules altérations des macro-polluants du SEQ-eau et à un seul indice biologique : l'I2M2. Notre méthode vise à faire une classification des motifs sur la base de leur support commun et de les associer à des stations pour lesquelles il y a eu un changement d'état de l'indice biologique ou pas.

1.1 Classification des motifs sur la base de leur support

Notre objectif est de proposer une classification des motifs en fonction de la similarité de leurs supports. Pour cela nous avons réalisé la démarche exposée ci-dessous.

A chaque extraction de motifs, l'algorithme PRESTOR fournit une série de résultats, parmi lesquels figurent les motifs, en fichiers images, identifiés par un

numéro allant de 0 à X, ainsi qu'un fichier « support.csv » (Tableau 11) listant pour chaque motif les codes des stations et des dates vérifiant ce motif (chapitre III, 2.4.4, Figure 20).

Tableau 11 : Exemple d'extrait de fichier « support » obtenu pour l'HER 18 (extraction HER 18, I2M2 ; 60 mois, fréquence minimale 0,7)

Identifiant motif	station	date
0	407147	17/08/2012
0	407171	20/07/2010
0	407171	03/08/2007
0	407171	06/08/2008
0	407171	15/08/2012
….		
182	418654	13/09/2010
182	418654	29/07/2008
182	418654	24/09/2007
182	418703	30/07/2008

Les stations et les dates, qui correspondent à une séquence, sont concaténées, puis le fichier « support.csv » est transformé en un tableau de contingence dont les individus sont les motifs, identifiés par leur numéro, et les variables sont les séquences (Tableau 12).

Tableau 12 : Exemple d'extrait de tableau de contingence obtenu avec les motifs en individus et les séquences (stations-dates) en variables (extraction HER18 ; I2M2 ; 60 mois, fréquence minimale 0,7)

Identifiant motif	407097-03/08/2007	407097-03/08/2011	407097-04/08/2008	…	418703-30/07/2008	418703-31/07/2012
0	0	0	0	…	0	0
1	0	0	0	…	0	0
2	0	0	0	…	0	0
3	0	0	0	…	0	0
4	0	0	0	…	0	0
…	…	…	…	…	…	…
179	0	0	0	…	1	0
180	0	0	0	…	1	0
181	0	0	0	…	1	0
182	0	0	0	…	0	0

La distance choisie pour la classification des motifs est le coefficient de similarité de Jaccard (Jaccard, 1901) dont la formule est donnée ci-dessous (Equation **5**).

$$I_{JACCARD} = \frac{a}{a + b + c}$$

Où :

- a est le nombre de séquences partagées par les motifs M1 et M2,
- b est le nombre de séquences possédées par le motif M1 et pas par le motif M2,
- c est le nombre de séquences possédées par le motif M2 et pas par le motif M1,
- d est le nombre de séquence non possédées par aucun des motifs M1 et M2.

Equation 5 : Indice de similarité de Jaccard (Jaccard, 1901)

Ce coefficient a l'avantage, par rapport à d'autres indices de similarité, de donner plus de poids au nombre de variables (ici les séquences) possédées (a, b, c) et partagées (c) par deux individus (ici les motifs) plutôt qu'au nombre de variables non possédées (d) par aucun des deux individus. Il donne donc plus de poids aux variables partagées (a), bien que leur nombre soit en général bien inférieur au nombre de celles non possédées (d).

La classification finale est réalisée par la méthode des K-means (Thorndike, 1953) sur la matrice de similarité calculée avec le coefficient de Jaccard. Le nombre de répétitions est fixé à 10. Le nombre de classes choisies est la racine carrée du nombre de motifs. Ces traitements statistiques ont été réalisés sous R-Studio (https://www.rstudio.com) en utilisant la librairie Ade-4 (Dray and Dufour, 2007). Le script utilisé est donné ci-dessous (Figure 53). Le Tableau **13** fournit un extrait des résultats obtenus : le classement des motifs par classe, ainsi que la taille du support vérifiant le motif.

```
#inclusion de la librairie utilisée
>library("ade4")
#lecture du tableau binaire
>donnees<-read.table(file = "support.csv", header = TRUE, sep = ";", row.names = 1)
#conversion du tableau en matrice
>dfmat<-as.data.frame(donnees)
#calcul de la distance de Jaccard entre les motifs
>dist_JACCARD<-dist.binary(dfmat, method = 1, diag = FALSE, upper = FALSE)
#algorithme des K-means pour la classification en x = sqrt(nombre de motifs) classes
>classes<-kmeans(dist_JACCARD, centers = x, iter.max = 10, nstart = 1, algorithm = "Hartigan-Wong ", trace = FALSE)
#récupération des résultats de la classification
>classes_num<-classes[1]

#transfert des données dans un fichier CSV
>write.table(classes_num, "classes-support.csv", sep = ";", col.names = TRUE)
```

Figure 53 : Script utilisé pour la classification des motifs sous R-Studio

Tableau 13 : Exemple d'extrait de tableau triant les motifs par classe et nombre de séquences vérifiant le motif (extraction HER18 ; I2M2 ; 60 mois, fréquence minimale 0,7)

Classe	motifs	Nombre de séquences
1	37	14
1	39	12
1	41	14
1	43	14
1	56	13
1	61	12
1	70	14
1	84	13
1	85	13
1	88	13
1	90	13
1	99	12
2	75	11
2	82	11
2	87	10
2	95	10
2	103	11
2	105	11
3	112	4
…		

Deux motifs sont sélectionnés par classe. Le premier critère de choix est basé sur la taille du support qui doit s'approcher de la médiane des tailles de supports constituant la classe. Ce critère permet de s'affranchir des motifs de support élevé et par conséquent peu informatifs et, à l'inverse, des motifs de faible support et donc trop complexes. Lorsque plus de deux motifs ont pour support la médiane, nous utilisons le maximum de la combinaison (P) des mesures d'intérêts : la fréquence du motif (F), émergence (E), la singularité (S) et la complexité (C). La formule de P est rappelée en Equation 6, ci-dessous. Enfin, dans le cas où, dans ces deux motifs sélectionnés selon les critères précédents, les items sont identiques, nous choisissons le motif suivant présentant au moins un item différent, lorsque c'est possible.

$$P = F \times C \times S + E$$

Où :

- F : fréquence du motif,
- C : complexité du motif
- S : singularité du motif,
- E : émergence du motif,
-

Equation 6 : Combinaison P des mesures d'intérêt pour un motif

1.2 Repérer un changement d'état biologique

Le programme PRESTOR propose plusieurs critères de choix pour l'extraction des motifs dont la sélection géographique des stations : sur toute la France ou pour une HER, et les contextes : tous les indices biologiques ou un seul d'entre eux. Pour l'instant, nous n'avons pas la possibilité de sélectionner des stations sur la base de l'évolution de leur état biologique. Aussi, l'identification de motifs caractérisant un changement d'état biologique ne peut se faire qu'à postériori, c'est-à-dire en sélectionnant les motifs extraits sur l'ensemble d'une HER qui sont vérifiés par des stations pour lesquelles une diminution ou, à l'inverse, une augmentation, de l'état d'un indice biologique donné a été observée. La requête ci-dessous (Figure 54) appliquée à la base de données FresQueau, nous a permis d'identifier les stations sur lesquelles la classe d'état de l'indice biologique invertébré I2M2 diminue ou, à l'inverse, augmente. La Figure 55 schématise l'ensemble de notre démarche.

```
SELECT r1.station_id, r1.date, r1.valeur, r2.date, r2.valeur

FROM hydrobiologie.parametre p1 JOIN hydrobiologie.resultat r1 ON p1.id=r1.parametre_id

JOIN hydrobiologie.resultat r2 ON r1.station_id=r2.station_id

JOIN hydrobiologie.parametre p2 ON p2.id=r2.parametre_id

WHERE r1.station_id IN (/* liste de toutes les stations-dates de la HER 18 */)

AND p1.id=70  /* 70 correspond à l'I2M2 dans la base de données */

AND p2.id=70

/* une hausse de la valeur de l'I2M2 correspond à une détérioration de celui-ci et inversement */

AND r1.valeur<r2.valeur /*pour une détérioration de l'I2M2 */

AND r1.date<r2.date
```

Figure 54 : requête SQL appliquée à la base de données FresQueau sur les résultats de l'I2M2, de l'HER18, permettant d'identifier les stations sur lesquelles la classe d'état de cet indice diminue ou, à l'inverse, augmente

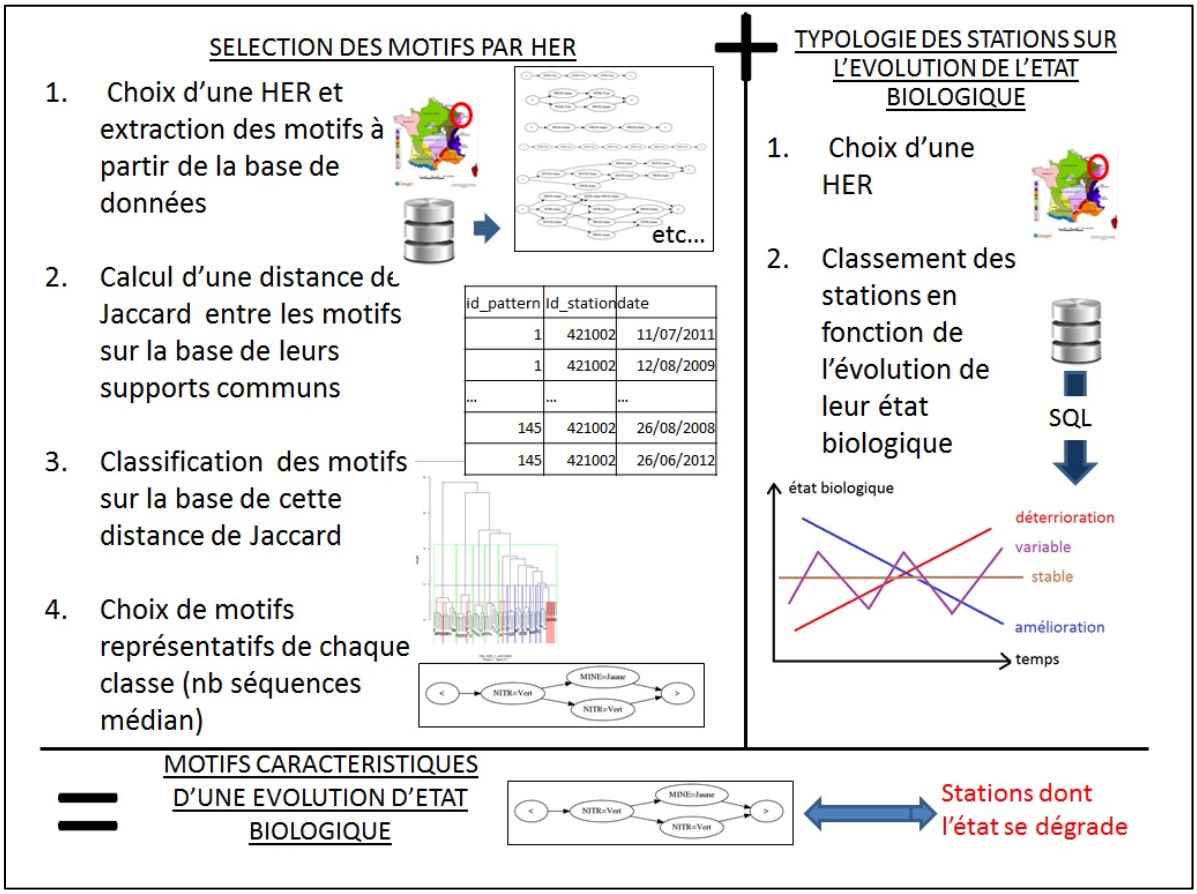


Figure 55 : Schéma de la démarche élaborée

2 Application à l'hydro-écorégion Alsace (HER 18)

Les classes d'état biologique ont été définies selon les seuils des grilles DCE de 2012 (MEEM, 2012). Cette HER comporte 30 stations. Le Tableau 14 indique le nombre de séquences et de stations disponibles par état de l'I2M2.

Tableau 14 : Nombre de séquences et de stations disponibles sur l'HER 18 par classe d'état de l'indice biologique I2M2

Etat de l'I2M2	Nombre de séquences	Nombre de stations
I2M2 Bleu	17	7
I2M2 Vert	42	16
I2M2 Jaune	36	18
I2M2 Orange	30	11
I2M2 Rouge	14	4

2.1 Typologie de l'état biologique I2M2 des stations de l'HER18

La requête SQL appliquée aux 30 stations permet de les classer en cinq types en fonction de l'évolution de leur état biologique I2M2 sur la période utilisée 2007-2013 Nous distinguons les stations stables ou peu variables, généralement d'état très bon à moyen : 6 stations, codées S(1-3), soit d'état moyen à mauvais : 4 stations, codées S(-3-15), et à l'opposé les stations d'état variable, soit dont l'état s'améliore (8 stations, codées V(+)), soit celles dont l'état se dégrade (7 stations, codées V(-)). Enfin, 1 station a un état très variable et est codée TV.

La Figure 56 représente graphiquement les évolutions des états biologiques pour les quatre premiers types proposés. Le Tableau 15 donne la liste des stations suivant cette typologie.

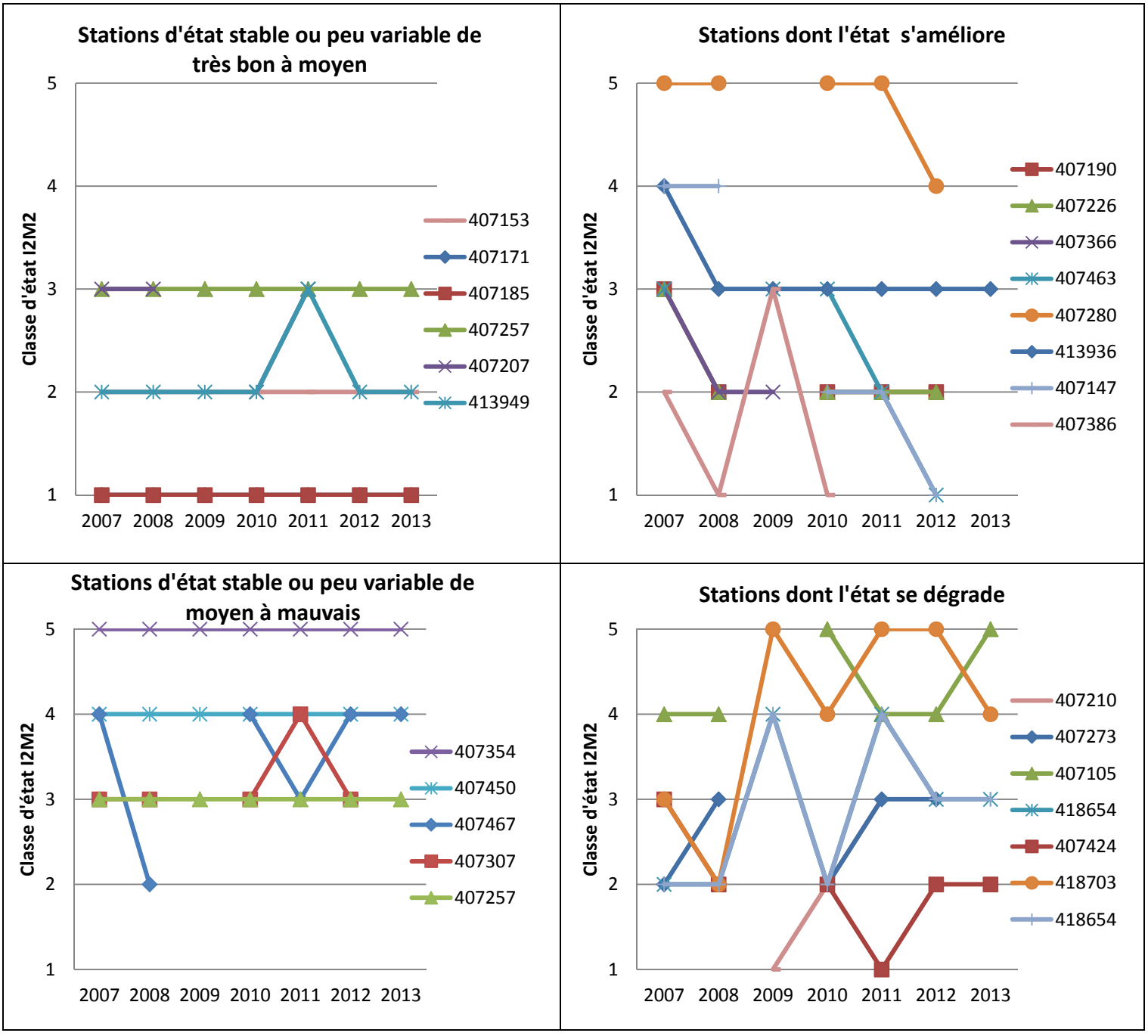


Figure 56 : Evolution interannuelle de l'état biologique de l'I2M2 pour les 26 stations utilisées dans l'extraction de motifs de l'HER18

Tableau 15 : Typologie de l'état biologique des stations de l'HER 18 utilisées pour les extractions de motifs

Type de stations	Code du type	Identifiants des stations	Nombre de stations
Etat stable ou peu variable de très bon à moyen	S(1-3)	407171, 407185, 407153, 407257, 407207, 413949,	6
Etat stable ou peu variable de moyen à mauvais	S(3-5)	407450, 407354, 407307, 407467	4
Etat s'améliorant globalement	V(+)	407147, 407190, 407226, 407280, 407366, 407386, 407463, 413936	8
Etat se dégradant globalement	V(-)	407105, 407210, 407273, 407424, 407448, 418654, 418703	7
Etat très variable	TV	407097	1

2.2 Extractions et classification des motifs de l'HER 18

Afin de pouvoir tester une longueur de séquences importante, nous nous limitons aux macro-polluants et avons enlevé les 3 altérations pour lesquelles les résultats sont tous de très bonne ou de bonne qualité : la température TEMP, l'acidité ACID, l'eutrophisation ou les effets de prolifération végétales EPRV. Nous travaillons donc avec six altérations du SEQ-eau sur neuf : les paramètres de minéralisation (MINE), les matières organiques et oxydables (MOOX), les matières azotées hors nitrates (AZOT), les nitrates (NITR), les matières phosphorées (PHOS), et les particules en suspension (PAES). Une première extraction de motifs réalisée pour une fréquence minimale de 0,93 et une longueur de séquences de 24 mois, nous a permis d'extraire 72 motifs, dont 89% se limitaient aux quatre items MINE_Bleu, MOOX_Vert, AZOT_Vert et PAES_Vert. Dans les autres, c'est-à-dire 11% de motifs, où d'autres items étaient présents, ces quatre items restaient dominants à 85%. Comme expliqué dans le chapitre précédent (chapitre III, 2.5.1), cette première

extraction nous permet de supprimer ces items dominants afin de rendre possible l'extraction d'autres motifs avec des items différents. Après suppression de ces quatre altérations dans les classes de qualité indiquées, nous avons pu réaliser une extraction pour une longueur de séquences de 5 ans (60 mois), soit la totalité de la période pour laquelle nous disposons de données physico-chimiques (2007-2012), pour une fréquence minimale de 0,7. Nous avons obtenu 183 motifs. Les caractéristiques de ces deux extractions sont résumées dans le Tableau 16 ci-dessous.

Tableau 16 : Caractéristiques des extractions de motifs réalisées sur l'HER 18

N° d'extrac-tion	HER	Période	Longueur de séquences (mois)	Fréquence minimale	Nombre de stations utilisées	Nombre de séquences utilisées	Nombre de motifs	Altérations supprimées
1	18	2007-2013	24	0,93	26	139	72	Micropolluants
2	18	2007-2013	60	0,7	26	139	183	Micropolluants & MINE Bleu, MOOX Vert, AZOT Vert, PAES Vert

Nous appliquons la classification décrite précédemment à l'extraction 2. Le nombre de classes choisies est 14 (racine carrée arrondie de 183). L'annexe 6 précise la répartition de l'ensemble des 183 motifs par classes, ainsi que leur support. L'annexe 7 liste les motifs sélectionnés par classe, sur la base de la médiane de leur support, et triés par ordre de leur combinaison P décroissante. Sont précisés les types et le nombre d'items présents par motif. La première sélection proposée des motifs par la classification sur la base de leur support, en conservant tous ceux dont le support est égal à la médiane, permet de limiter le nombre de motifs de l'extraction de 183 à 72, soit une diminution de 61%. Parmi ces motifs sélectionnés, 45 sont dominants (25% des motifs de départ) : leur contexte de génération est le contexte dominant, c'est-à-dire celui dont la fréquence est la plus élevée. Leur niveau de singularité est élevé : en effet, seuls 6 d'entre eux sont

communs à plusieurs contextes de génération. En général, leur combinaison maximale P est supérieure à 1 – il varie de 1,11 pour la classe 14 à 12,14 pour la classe 7 – excepté pour les classes 1, 3, 4 et 5, pour lesquelles P maximum est inférieur à 0,02.

Deux motifs caractéristiques de chaque classe, sont sélectionnés selon trois critères : (i) avoir un support égal ou proche de la médiane des supports des motifs de la classe, (ii) avoir la combinaison P maximale, (iii) avoir au moins un item différent lorsque c'est possible (Tableau 18).

Il est rappelé que les altérations MINE_Bleu, MOOX_Vert, AZOT_Vert et PAES_Vert étaient dominantes à 89% dans les motifs de la 1ère extraction et qu'ils ont été retirés pour obtenir la 2ème extraction. Nous pouvons estimer que les motifs obtenus lors de la 2ème extraction devraient être complétés chacun de la succession d'environ 50 itemsets – 89 % des 60 mesures physico-chimiques dont nous disposons en moyenne par station – contenant les quatre items MINE_Bleu, MOOX_Vert, AZOT_Vert et PAES_Vert.

2.3 Correspondance des classes des motifs et des typologies des stations les vérifiant

Pour chaque classe de motifs obtenue, nous dénombrons le nombre de stations réparties par type d'état biologique et vérifiant les motifs (annexe 8). La médiane, les premier et troisième quartiles du nombre de stations sont calculés par type de stations et sur l'ensemble des classes. Lorsque le nombre de stations par classe est inférieur au premier quartile, le type est considéré comme faiblement présent dans la classe. Lorsque ce nombre est supérieur au troisième quartile, le type est considéré comme fortement présent dans la classe. Enfin, lorsque ce nombre est compris entre le premier et le troisième quartile, le type est considéré comme moyennement présent dans la classe.

Tableau 17 : Caractérisations des classes des motifs obtenues pour l'extraction 2 (HER18, I2M2, 60 mois, fréquence minimale 0,7), de la typologie de leurs stations et de leur tendance

- Légende des polices du code des types de stations : normal = type faiblement présent, **gras** = type moyennement présent, **gras et police 14** = type fortement présent
- Légende des tendances des états biologique : ~ variables ; - - dégradation et/ou stable en état dégradé ; + + amélioration et/ou stable en état très bon à moyen

N° de classe	Nombre de motifs	Médiane des supports	Codes types stations vérifiant les motifs de la classe	Tendance
1	6	10	S(1-3) ; V(+) ; S(3-5) ; V(-) ; TV	~
2	7	3	**S(1-3)** ; **V(+)** ; S(3-5) ; V(-)	+ +
3	7	9	**S(1-3)** ; **S(3-5)** ; **V(+)** ; **V(-)** ; TV	~
4	7	6	**S(1-3)** ; **V(+)** ; S(3-5) ; V(-) ; **TV**	~
5	4	23	S(1-3) ; V(+) ; S(3-5) ; V(-)	~
6	2	5,5	S(1-3) ; V(+)	+ +
7	63	4	**S(3-5)** ; **V(-)**	- -
8	48	12	S(1-3) ; V(+) ; **S(3-5)** ; **V(-)** ; TV	- -
9	8	9	S(1-3) ; V(+) ; **S(3-5)** ; V(-) ; **TV**	- -
10	6	4	**S(1-3)** ; **V(+)** ; V(-)	+ +
11	2	5,5	S(1-3) ; **V(+)** ; V(-)	+ +
12	2	5,5	S(1-3) ; **V(+)** ; V(-)	+ +
13	4	4	S(1-3) ; **V(+)** ; V(-)	+ +
14	17	12	V(+) ; **S(3-5)** ; **V(-)** ; **TV**	- -

Le Tableau 17 indique le nombre de motifs obtenus par classe, la médiane de leur support, ainsi que la typologie des stations vérifiant ces motifs. Est également indiquée dans ce tableau la tendance générale d'évolution de l'état biologique I2M2 pour chaque classe. Cette tendance est définie en fonction des types de stations dominants dans la classe de motifs. Trois tendances sont proposées :

- un état biologique variable qui regroupe les quatre classes : 1, 3, 4 et 5 ; il est à noter qu'il s'agit des classes pour lesquelles nous avions observé, par ailleurs, une très faible combinaison P ;
- un état biologique s'améliorant et/ou restant stable entre les états très bon à moyen, qui regroupe les six classes 2, 6, 10, 11, 12 et 13 ;
- un état biologique se dégradant et/ou restant stable entre les états moyen et mauvais, qui regroupe les quatre classes 7, 8, 9 et 14.

Nous désignons respectivement ces trois groupes par Groupe V, Groupe A et Groupe D. Le Tableau 18 indique la répartition des motifs caractéristiques par classe dans ces trois groupes.

Les figures 57, 58 et 59 représentent les motifs caractéristiques respectivement de chacun de ces groupes. La classe 7 est caractérisée par deux très grands motifs : le 165, qui compte 139 items, et le 171, qui est le plus grand obtenu avec 245 items. Ces deux motifs sont vérifiés par quatre stations. Ils ont été extraits pour le contexte I2M2 rouge. Le motif 171 compte notamment 77 items PHOS_Jaune, ce qui peut paraitre beaucoup pour des stations pour lesquelles les analyses physico-chimiques ont été mesurées avec une fréquence généralement mensuelle, pour la période 2007-2012, soit environ 60 mesures. Après vérifications, le nombre de mesures disponibles est plus élevé sur ces stations : la station 497354, notamment, possède 79 résultats sur les paramètres du phosphore. Le motif 165 est représenté sur la Figure 57 et, de façon plus lisible, en annexe 9.

Tableau 18 : Choix de 2 motifs caractéristiques par classe regroupée par tendance de l'état biologique

Classe	Motif	Nb supports	Contexte de génération	Contexte dominant	Items du motif
GROUPE V : classes de motifs d'état biologique variable					
1	33	10	I2M2=Jaune	I2M2=Bleu	3 NITR_Vert
1	31	10	I2M2=Jaune	I2M2=Bleu	2 NITR_Vert
3	34	9	I2M2=Jaune	I2M2=Bleu	2NITR_Vert, 2 PAES_Jaune
3	36	9	I2M2=Jaune	I2M2=Bleu	3 NITR_Vert, 1 PHOS_Jaune
4	180	6	I2M2=Vert	I2M2=Rouge	3 PHOS_Jaune
4	177	6	I2M2=Vert	I2M2=Orange	2 PHOS_Jaune
5	182	23	I2M2=Vert	I2M2=Bleu	7 NITR_Vert
5	181	23	I2M2=Vert	I2M2=Bleu	5 NITR_Vert
GROUPE A : classes de motifs d'état biologique s'améliorant et/ou se maintenant entre très bon et moyen					
2	10	3	I2M2=Bleu	I2M2=Bleu	10 NITR_Vert, 2 MINE_Jaune
2	7	3	I2M2=Bleu	I2M2=Bleu	5 NITR_Vert, 3 MINE_Jaune
6	23	5	I2M2=Bleu	I2M2=Bleu	29 NITR_Vert, 9 MINE_Vert
6	3	6	I2M2=Bleu	I2M2=Rouge	1 MOOX_Jaune, 1 PHOS_Jaune
10	21	4	I2M2=Bleu	I2M2=Bleu	24 NITR_Vert, 11 MINE_Vert
10	17	4	I2M2=Bleu	I2M2=Bleu	17 NITR_Vert, 7 MINE_Vert
11	12	5	I2M2=Bleu	I2M2=Bleu	10 NITR_Vert, 2 PHOS_Jaune
11	4	6	I2M2=Bleu	I2M2=Orange	2 PHOS_Jaune
12	9	5	I2M2=Bleu	I2M2=Bleu	6 NITR_Vert, 1 PHOS_Jaune
12	2	6	I2M2=Bleu	I2M2=Rouge	1 PHOS_Jaune
13	22	4	I2M2=Bleu	I2M2=Bleu	20 NITR_Vert, 1MINE_Vert
13	1	4	I2M2=Bleu	I2M2=Rouge	1 NITR_Jaune
GROUPE D : classes de motifs d'état biologique se dégradant et/ou se maintenant dégradé					
7	165	4	I2M2=Rouge	I2M2=Rouge	8 MINE_Vert, 15 NITR_Jaune, 13 NITR_Orange, 43 PHOS_Jaune, 17 AZOT_Jaune, 27 MOOX_Jaune, 12 PAES_Jaune, 4 PAES_Rouge
7	171	4	I2M2=Rouge	I2M2=Rouge	10 MINE_Vert, 20 NITR_Jaune, 28 NITR_Orange, 77 PHOS_Jaune, 27 AZOT_Jaune, 48 MOOX_Jaune, 26 PAES_Jaune, 9 PAES_Rouge
8	74	12	I2M2=Orange	I2M2=Orange	2 NITR_Vert, 2 PHOS_Jaune, 1PAES_Rouge
8	101	12	I2M2=Orange	I2M2=Orange	3 NITR_Jaune, 5 PHOS_Jaune, 3 MOOX_Jaune
9	48	9	I2M2=Orange	I2M2=Orange	1 MOOX Bleu, 1 PHOS_Jaune
9	47	9	I2M2=Orange	I2M2=Orange	1 MOOX Bleu, 1 NITR_Jaune
14	94	12	I2M2=Orange	I2M2=Orange	3 PHOS_Jaune, 4 MOOX_Jaune
14	104	12	I2M2=Orange	I2M2=Orange	4 NITR_Jaune, 5 PHOS_Jaune, 1 MOOX_Jaune

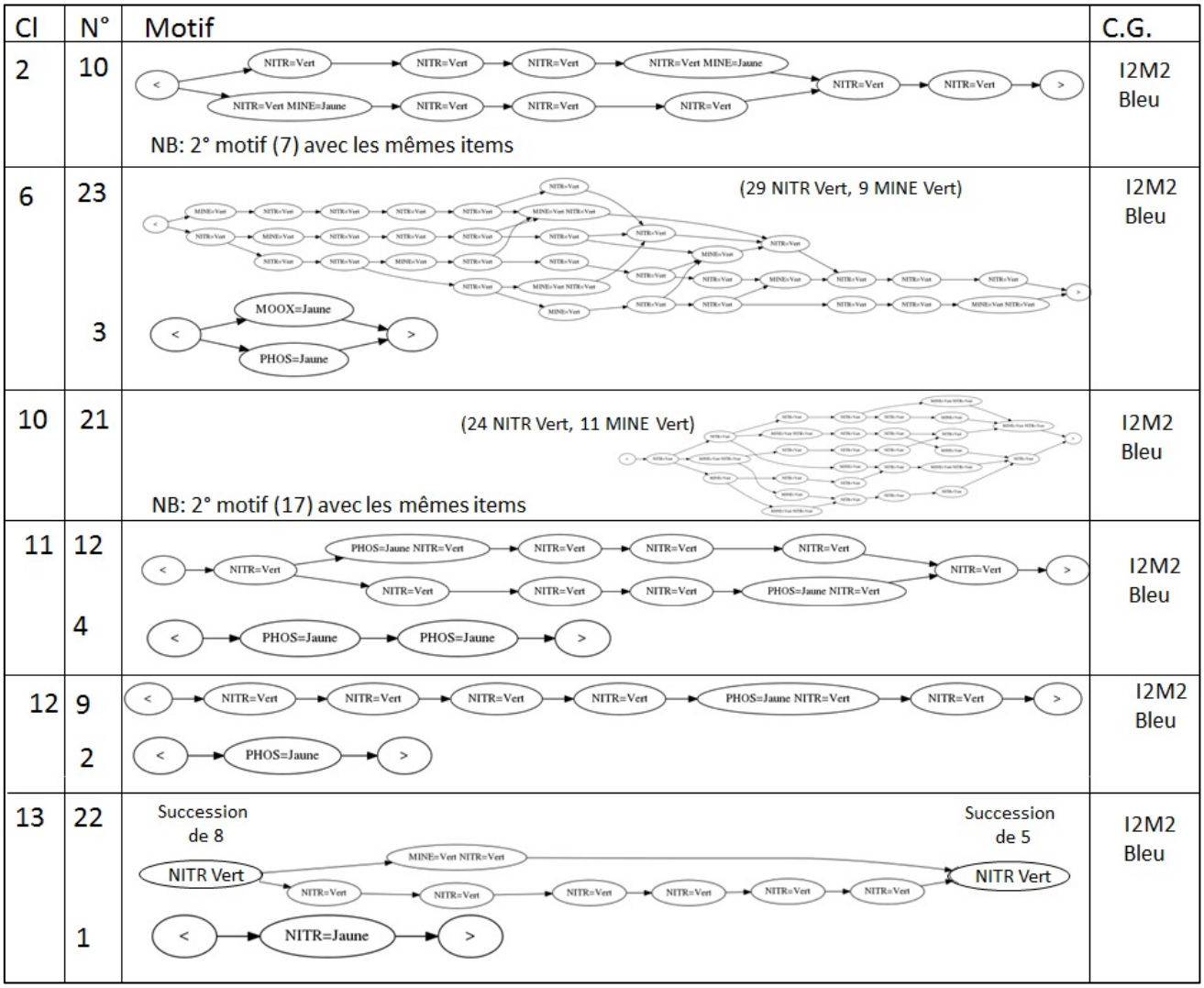


Figure 57 : Motifs caractéristiques des classes du groupe A pour lesquelles l'état biologique des stations s'améliore et/ou se maintient entre très bon à moyen (C.G. = Contexte de Génération des motifs)

Le groupe A a trois types de motifs caractéristiques. Dans les classes 6, 10 et 13, les premiers motifs caractéristiques ont de longues successions d'itemsets de bonne qualité : de 20 (motif 22, classe 13) à 35 itemsets (motif 23, classe 6), pour les deux altérations NITR_Vert ou MINE_Vert. Le nombre d'items NITR_Vert varie alors de 19 (motif 22, classe 13) à 29 (motif 23, classe 6) ; celui de MINE_Vert de 1 (motif 22, classe 13) à 11 (motif 21, classe 10). Dans les classes 2, 11 et 12, les premiers motifs ont des successions d'items plus courtes dominées par la seule altération de bonne qualité NITR_Vert – de 6 items (motif 9, classe 12) à 10 (motif 10, classe 2 et motif 12, classe 11) – associée à une altération de qualité moyenne n'apparaissant qu'une ou deux fois. Cette altération moyenne peut être MINE_Jaune (motif 10,

classe 2) ou PHOS_Jaune (motif 10, classe 2 et motif 12, classe 11). Enfin, dans les classes 6, 11 et 13, le deuxième motif caractéristique est court : limité à 1 ou deux items de qualité moyenne. Ces altérations moyennes peuvent être MOOX_Jaune (motif 3, classe 6), PHOS_Jaune (motif 3, classe 6 et motif 4, classe 11) ou NITR_Jaune (motif 1, classe 13).

Les motifs de ce groupe A comportent quatre altérations sur les six utilisées : MINE, MOOX, NITR, PHOS dans seulement deux classes de qualité : bonne et moyenne représentées par les couleurs verte et jaune. Les contextes de génération de ce groupe est l'I2M2 Bleu.

Le groupe D, celui où la qualité des stations se dégrade ou reste stable dans un état plutôt mauvais, a quatre types de motifs caractéristiques. Dans la classe 7, les motifs sont très grands avec par exemple 139 items pour le motif 165 et 245 pour le motif 171, qui est le plus grand de l'extraction. Les six altérations utilisées y sont présentes dans des qualités moyenne à mauvaise excepté pour la minéralisation qui n'apparait que de bonne qualité : MINE_Vert. Les nitrates apparaissent de qualité moyenne ou médiocre : NITR_Jaune ou NITR_Orange ; les particules en suspension de qualité moyenne ou mauvaise : PAES_Jaune ou PAES_Rouge ; les matières organiques, les matières azotées hors nitrates et les matières phosphorées de qualité moyenne : MOOX_Jaune, AZOT_Jaune, PHOS_Jaune. Les classes 8 et 14 ont des motifs de taille moyenne : les motifs 94 et 104 (classe 14), 101 (classe 8) ont entre 7 à 10 itemsets avec deux ou trois altérations toutes de qualité moyenne : MOOX_Jaune, PHOS_Jaune (motif 94) auxquels s'ajoute NITR_Jaune (motifs 104 et 101). La classe 8 a un premier motif caractéristique plus court : le motif 74, avec cinq itemsets, trois altérations différentes, mais avec trois niveaux de qualité allant de bonne à mauvaise : NITR_Vert, PHOS_Jaune et PAES_Rouge. Les motifs de la classe 9 sont limités à deux itemsets et deux items : le motif 48 a un PHOS_Jaune, le motif 47 a un NITR_Jaune et ils ont chacun un MOOX bleu.

Les motifs du groupe D comportent les six altérations utilisées : NITR, MINE, MOOX, AZOT, NITR, PHOS et PAES. Il comporte également toutes les classes de qualité, de très bonne à mauvaise. Les contextes de génération de ce groupe sont tous I2M2 Orange (classes 8, 9, 14) ou I2M2 Rouge (classe 7).

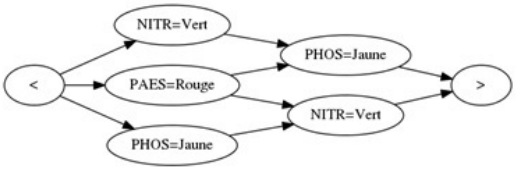
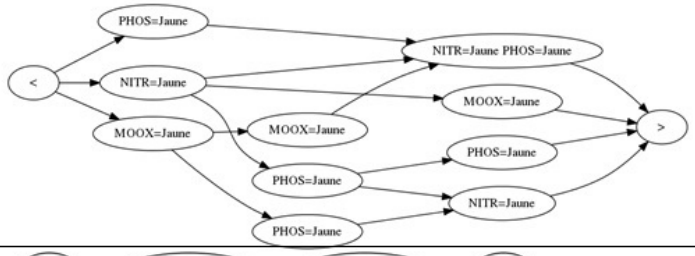
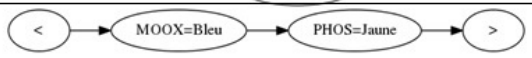
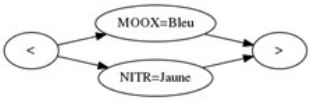
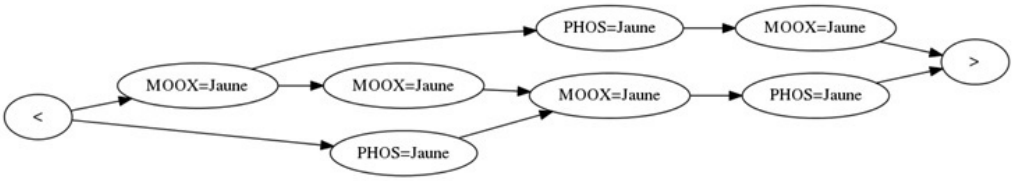
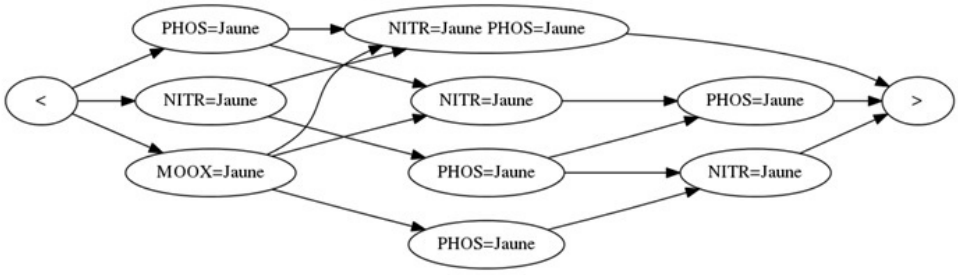
Cl	N°	Motif	C.G.
7	165	 245 items (8 MINE Vert, 15 NITR Jaune, 13 NITTR Orange, 43 PHOS Jaune, 17 AZOT Jaune, 27 MOOX Jaune, 12 PAES Jaune, 4 PAES Rouge) NB: 2° motif (171) avec les mêmes items	I2M2 Rouge
8	74		I2M2 Orange
	101		
9	48		I2M2 Orange
	47		
14	94		I2M2 Orange
	104		

Figure 58 : Motifs caractéristiques des classes du groupe D: pour lesquelles l'état biologique des stations se dégrade et/ou se maintient dégradé (le motif 165 est en grande taille en annexe 9) (C.G. = Contexte de Génération des motifs)

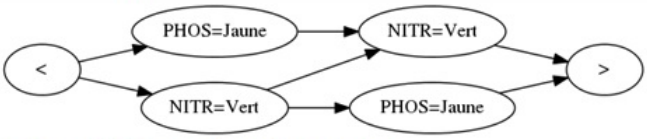
Cl	N°	Motif	C.G.
1	33	 NB: 2° motif (31) avec les mêmes items	I2M2 Bleu
3	34	 NB: 2° motif (36) avec les mêmes items	I2M2 Jaune
4	180	 NB: 2° motif (177) avec les mêmes items	I2M2 Vert
5	182	 NB: 2° motif (181) avec les mêmes items	I2M2 Vert

Figure 59 : Motifs caractéristiques des classes du groupe V : pour lesquelles l'état biologique des stations est variable (C.G. = Contexte de Génération des motifs)

Les motifs du groupe V se rapprochent de ceux du groupe A, mais sont en général plus courts. Les classes 1 et 5 sont limitées à des successions de trois à sept NITR_Vert, respectivement pour les motifs 33 et 182. La classe 4 est limitée à une succession maximale de trois PHOS_Jaune (motif 180). La classe 3 est limitée à quatre items maximum : deux PHOS_Jaune et NITR_Vert (motif 34).

Les motifs de ce groupe comportent seulement deux altérations sur les six utilisées : NITR et PHOS, dans seulement deux classes de qualité : bonne pour NITR_Vert et moyenne pour PHOS_Jaune. Les contextes de génération de ce groupe sont tous I2M2 Bleu (classe 1), I2M2 Vert (classes 4 et 5) ou I2M2 Jaune (classe 3).

3 Comparaison des classements de la qualité des stations et synthèse

Peut-on répondre à la question : existe-t-il des successions d'altérations physico-chimiques caractéristiques d'un changement d'état biologique à l'échelle spatiale d'une hydro-écorégion (HER) ?

Dans notre cas d'application : l'HER 18, pour les stations dont l'état biologique se dégrade ou n'est pas bon ou se dégrade, les motifs sélectionnés lorsque l'état biologique le plus dégradé est médiocre (I2M2 Orange, classes 8 et 14), comportent de 5 à 11 items d'altérations dégradées pour les matières organiques (motifs 94 et 104, classe 14), et/ou les nitrates (motif 94, classe 14 ; motifs 74 et 101, classes 8) e/out les matières phosphorées (motifs 74, 101, classe 8 ; motifs 94 et 104, classe 14) de qualité moyenne et/ou les particules en suspension de mauvaise qualité (motif 74, classe 8 . Les motifs des classes 8 et 14 caractérisent donc des stations assez polluées. Pour les quatre stations dont le pire état biologique est mauvais (I2M2 Rouge, classe 7), les motifs sélectionnés sont très grands : de 139 à 245 items (motifs 165 et 171, classe 7), et indiquent des pollutions chroniques pour cinq des six altérations utilisées : les matières organiques, phosphorées et azotées hors nitrates, de qualité moyenne, les nitrates de qualité moyenne ou médiocre et les particules en suspension, de qualité moyenne ou mauvaise. Pour une moyenne de 72 analyses disponibles sur la période 2007-2012, le motif 171 indique que 100 % des analyses de matières phosphorées n'étaient pas de bonne qualité, 39 % de celles des nitrates, 38 % de celles des matières organiques, et plus de 20% pour les particules en suspension et les matières azotées hors nitrates. Les motifs de la classe 7 caractérisent donc des stations polluées de façon chronique. Enfin les motifs de la classe 9 sont peu informatifs et caractérisent des stations de pollution variable.

Pour les stations dont l'état biologique s'améliore ou se maintient en bon état, les motifs sélectionnés (classes 2, 6, 10, 11, 12, 13) ont atteint au moins une fois le très bon état biologique (contexte de génération I2M2 Bleu). Globalement, il s'agit de stations pour lesquelles les nitrates et la minéralisation se maintiennent en bonne qualité. Pour une moyenne de 72 analyses disponibles sur la période 2007-2012, le motif 23 (classe 6) indique que 40 % des analyses de nitrates étaient de bonne qualité ; pour le motif 21 (classe 10) 15 % des mesures de minéralisation l'étaient également. Les motifs de plus de deux items des classes 2, 6, 10, 13 caractérisent donc des stations préservées. Par contre des pollutions ponctuelles, de qualité moyenne, sont relevées sur ces stations pour les altérations minéralisation (motifs 10, classe 1), matières organiques (motif 3, classe 6), nitrates (motif 1, classe 13) et matières phosphorées (motifs 3, classe 6 ; motif 12 et 4, classe 11 ; motif 9, classe 12). Ces altérations ne représentent pas plus de 2 items par motif, ce qui représente

environ 3% des données disponibles. Les motifs des classes 11 et 12 et ceux d'un à deux items des classes 3, 4 et 13 caractérisent donc des stations peu polluées n'ayant pas perdu leur capacité de résilience.

Pour les stations dont l'état biologique est variable, les motifs sélectionnés des classes 1, 3, 4 et 5 sont peu informatifs et caractérisent des stations de pollution variable.

Suite à l'analyse de l'évolution de l'état biologique défini par I2M2, nous avions proposé un premier classement des stations de l'HER18 en cinq types :

- six stations, codées S(1-3), en état stable ou peu variable de très bon à moyen,
- quatre stations, codées S(3-5), en état stable ou peu variable de moyen à mauvais,
- huit stations, codées V(+) dont l'état s'améliore,
- sept stations, codées V(-) dont l'état se dégrade,
- une station, d'état très variable.

L'analyse des motifs caractéristiques des 14 classes, issues de la classification réalisée sur la base des supports des motifs, nous permet à présent de proposer un nouveau classement des stations de l'HER18. Ce classement correspond à l'évolution des pollutions de l'eau par les macro-polluants subies par les stations. Nous distinguons :

- deux stations préservées,
- cinq stations peu polluées et résilientes,
- quatre stations assez polluées,
- quatre stations polluées de façon chronique,
- 11 stations avec des niveaux de pollution variables sur la période de mesures disponibles.

Le Tableau 19 propose le croisement des deux types de classements des stations.

Ainsi, parmi les stations dont l'état biologique I2M2 s'est amélioré, trois stations sont peu polluées et résilientes : les stations 407386, 407147, 407463. Elles ont subi des pollutions ponctuelles, de qualité moyenne, pour les matières organiques et les matières phosphorées, auxquelles s'ajoutent l'altération minéralisation pour la première et l'altération nitrates pour les deux autres, également de qualité moyenne. L'amélioration de leur état biologique devrait se maintenir, sous réserve qu'elles soient préservées de nouvelles pollutions. La station 407280 est polluée de façon chronique par toutes les altérations utilisées exceptée la minéralisation ; son amélioration biologique n'est sûrement que ponctuelle. Les quatre stations 407226, 407190, 407366, 413936 ont des niveaux de pollution mal définis. Prédire la poursuite de l'amélioration biologique observée est difficile.

Tableau 19 : Classement des stations de l'HER18 en fonction de l'évolution de leur état biologique I2M2 et de l'évolution de la pollution de leur eau, définies sur la base des motifs caractéristiques

STATIONS	CLASSEMENT DES STATIONS D'APRES L'EVOLUTION DE L'ETAT BIOLOGIQUE					CLASSEMENT DES STATIONS D'APRES LES MOTIFS CARACTERISTIQUES				
	S (1-3)	V(+)	S(3-5)	V(-)	TV	Préser-vées	Peu polluées & résilientes	Assez polluées	Polluées chroniques	Varia-bles
407171	X					X				
407185	X					X				
407153	X									X
407257	X									X
407207	X									X
413949	X									X
407386		X					X			
407147		X					X			
407463		X					X			
407280		X							X	
407226		X								X
407190		X								X
407366		X								X
413936		X								X
407467			X					X		
407450			X					X		
407354			X						X	
407307			X							X
407210				X			X			
407424				X			X			
418654				X				X		
407105				X					X	
418703				X					X	
407273				X						X
407448				X						X
407097					X			X		

Pour les stations dont l'état biologique I2M2 s'est dégradé, deux stations sont peu polluées et résilientes : les stations 407210 et 407424. Elles ont subi des pollutions ponctuelles, de qualité moyenne, pour les matières organiques, les matières phosphorées et les nitrates. Leur dégradation pourrait être stoppée si elles ne subissent pas d'autres pollutions. A l'opposé, les deux stations 407105 et 418703 subissent des pollutions chroniques pour toutes les altérations utilisées excepté la minéralisation, leur dégradation sera difficile à enrayer. La station 418654 subit des pollutions assez répétées par les matières organiques, matières phosphorées, nitrates de qualité moyenne et les particules en suspension de mauvaise qualité. Des efforts importants seraient nécessaires pour enrayer sa dégradation. Les deux stations 407273 et 407448 ont des niveaux de pollution mal définies. Prédire la poursuite de la dégradation biologique observée est difficile.

En conclusion, la méthode proposée de classification des motifs sur la base de leur support permet de limiter le nombre de motifs à analyser de façon significative. Dans le cas traité, les motifs sélectionnés ont permis de classer efficacement les stations en fonction de l'évolution des pollutions de l'eau par les macro-polluants et de répondre à la question posée.

Pour l'instant, le programme PRESTOR d'extraction des motifs ne permet pas de sélectionner des lots de stations pour lesquelles un état biologique s'améliore ou se détériore. La sélection des stations ne peut se faire que sur le critère géographique national ou d'une HER donnée. C'est pourquoi nous avons dû extraire tous les motifs de l'HER 18 et définir par ailleurs les tendances d'évolution des états biologiques. De plus, pour pouvoir extraire des motifs sur une longueur de séquences égale à la période disponible (60 mois, 2007-2012), nous avons aussi dû limiter les altérations à traiter et ne prenant que les macro-polluants. Ajouter au programme PRESTOR la possibilité de sélectionner un lot de stations, par HER, sur la base de l'évolution d'un état biologique, défini par un indice donné, et limiter l'extraction des motifs à cette seule sélection, devrait permettre d'augmenter le nombre d'altérations utilisables et d'obtenir des motifs spécifiques plus pertinents. Nous pourrions alors profiter des spécificités de chaque indice biologique dépendant des différences de réponses de chaque groupe taxonomique (Lafont, 2001).

Chapitre VI : Synthèse, discussion et perspectives

L'ambitieux objectif de la Directive Cadre européenne sur l'Eau (DCE, (European Council, 2000) vise à préserver et restaurer le bon état écologique de toutes les masses d'eau. Les difficultés de sa mise en œuvre concernent ses trois étapes : l'évaluation de l'état écologique, l'identification des mesures de restauration compte tenu de la capacité de résilience des masses d'eau à restaurer, enfin, l'évaluation des effets des restaurations. L'évaluation des pressions en cause dans la non atteinte du bon état est particulièrement difficile dans les contextes, largement répandus, de pressions multiples (Reyjol et al., 2014). Par ailleurs, la DCE a engendré dans plusieurs pays un renforcement des efforts de surveillance et donc de production de données. Sur les rivières, ces données sont à présent des données massives et complexes sur lesquelles peuvent s'appliquer des méthodes de fouilles issues du domaine informatique.

Nous avons testé ici l'application des méthodes de fouilles de données non supervisées pour répondre à une partie des questions posées par la DCE. En particulier, nous proposons l'extraction de motifs temporels partiellement ordonnés à partir des séquences de pressions physico-chimiques qui précèdent un état biologique, dans l'objectif d'identifier les pressions qui ont conduit à cet état. A l'échelle de la France métropolitaine, nous avons exploré les motifs obtenus à la fois sur différents groupes biologiques, pour différentes longueurs de séquences, et en utilisant différentes méthodes d'agrégation et de discrétisation des données. Nous avons travaillé sur quatre groupes biologiques et cinq indices s'y rapportant : les deux indices IBGN (AFNOR, 2004a) et I2M2 (Mondy et al., 2012) basés sur les macro-invertébrés, l'IPR (AFNOR, 2004b) basé sur les poissons, l'IBD (AFNOR, 2007) basé sur les diatomées, l'IBMR (AFNOR, 2003) basé sur les macrophytes. Les résultats des indices biologiques étaient discrétisés en état biologiques d'après les grilles DCE (MEEM, 2012). Nous avons testé cinq longueurs de séquences de 3, 6, 12, 18 et 24 mois. Enfin nous avons agrégés et discrétisés les résultats physico-chimiques dont nous disposions sur l'eau en altérations réparties en cinq classes de qualité ou d'état grâce aux grilles du SEQ-eau (MEDD et AE, 2003) et les mêmes

grilles DCE que pour les indices biologiques. Pour pouvoir extraire des motifs nous avons développé à la fois une base de données relationnelle et l'algorithme PRESTOR, s'appuyant sur la méthode de fouille élaborée par Fabrègue et al. (2014). L'extraction de motifs à grande échelle sur le territoire français métropolitain limite leur pertinence et les interprétations qui peuvent en être faites. Aussi proposons-nous d'appliquer cette méthode à l'échelle des hydro-éco-régions (HER) (J. Wasson et al., 2004).

Nous discuterons ici des apports, des limites et des perspectives de ce travail autour de trois thèmes :

- l'application des méthodes de fouille de données aux données de rivières,
- le programme d'extraction de motifs PRESTOR,
- les motifs temporels extraits aux échelles nationale et d'une HER.

1 La fouille non supervisée appliquée aux données rivières : apports, limites et perspectives

Appliquer la fouille non supervisée aux données rivières nécessite plusieurs pré-requis, dont le principal est d'avoir des données conséquentes, et de mettre en place un travail pluridisciplinaire entre thématiciens et informaticiens.

1.1 Les pré-requis indispensables : avoir des données structurées en base de données relationnelle

Le premier pré-requis pour faire de la fouille de données est d'avoir des données ! Même si les méthodes de fouille se sont révélées au grand public en même temps que la production de données massives (« *big data* »), comme évoqué en introduction (chapitre I, 3.1), elles peuvent être appliquées à de petits jeux de données. C'est ce que nous avons fait avec l'analyse relationnelle de concepts sur le jeu de données de la plaine d'Alsace, composé d'une quarantaine de stations (chapitre II, 2.6.2).

Sur les données de surveillance des rivières, nous avons évoqué la difficulté d'accéder notamment aux listes floristiques et faunistiques malgré les efforts constants des services de l'Etat pour la bancarisation et la mise à disposition du public de ces données (chapitre I, 2.3).

Les données ne suffisent pas, avoir leur méta-données est également nécessaire (Gibert et al., 2018), ainsi que maitriser leur qualité comme évoqué en introduction (chapitre I, 3.2). Des indicateurs de confiance peuvent être associés aux résultats de fouille (Berrahou et al., 2015) et permettre de pondérer la prise en compte des données en fonction de leur niveau de qualité. Sur les données nationales utilisées pour l'extraction des motifs (chapitres III, IV et V), nous avons dû mener un travail important de nettoyage des données, ce qui correspond aux étapes de vérification des indicateurs de qualité: cohérence, unicité, exactitude et fraicheur (Berti-Equille, 2004), avant de les intégrer dans la base de données créée pour les recevoir. Pour ne citer que quelques exemples : les systèmes de projection des coordonnées géographiques des stations de surveillance ont dû être homogénéisés ; les coordonnées parfois corrigées comme dans le cas de la Corse où les abscisses et les ordonnées des stations de surveillance étaient inversées : le Cap Corse était orienté vers l'Est ! Malgré le SANDRE, pour les paramètres physico-chimiques, un travail important de vérification et d'homogénéisation a été réalisé sur les codes des paramètres, leur unités de mesures (des nitrates mesurés en « mg/L » ou en « mg/L de nitrates » ne seront pas considéré comme deux résultats comparables).

Enfin, ces données doivent être structurées en base de données relationnelles, afin de maitriser et pouvoir explorer les liens établis entre elles. Ainsi l'analyse relationnelle de concepts, appliquée aux données de la plaine d'Alsace que nous avions produites précédemment (Grac et al., 2006), cherche des régularités entre les stations de mesures, les mesures de leurs paramètres physico-chimiques, les taxons qui y ont été échantillonnés et les modalités des traits biologiques et écologiques de ces taxons (Chapitre II, 2.6.2). Une des raisons pour lesquelles nous n'avons pas testé l'extension de cette méthode aux données de toute la France est que, dans les données mises à notre disposition, le lien entre les listes taxonomiques

et les stations de mesures n'était pas correctement établi et nécessitait à nouveau un travail important de mise en relation des données.

Ainsi avoir des données reliées entre elles et de qualité suffisante, la condition de départ pour appliquer des méthodes de fouille, n'est pas simple. C'est une étape chronophage qui nécessite la collaboration d'informaticiens spécialisés en base de données et de thématiciens, dans un permanent aller-retour d'identification d'un problème, de recherche de sa cause et de mise en œuvre d'une solution, avant de passer au problème suivant. Même si tous les projets de recherche sur les hydrosystèmes doivent inclure cette étape, et qu'une certaine concurrence existe pour l'obtention de données en quantité et qualité suffisantes, elle est rarement abordée car peu vendeuse dans les publications, mises à part celles faites en sciences de la donnée. Ce type de publications a, jusqu'à présent, peu concerné les données sur l'eau mis à part nos propres travaux (Berrahou et al., 2015; Bimonte et al., 2015).

1.2 Expérience menée en fouille non supervisée : de la richesse et de la difficulté du travail pluridisciplinaire

Nous avons expliqué que la fouille de données n'est qu'une étape d'un processus appelé extraction de connaissances de base de données (ECDB) (Fayyad et al., 1996) (chapitre I, 3.2). La collaboration entre informaticiens et thématiciens, ici les hydroécologues, ne se limite pas à la simple étape de préparation des données. Elle doit commencer avant, dès la conception de la question posée. C'est de la bonne compréhension entre thématiciens et informaticiens que dépendra la transformation de la question posée en question formelle, utilisable par les méthodes de fouille. Cette collaboration doit se prolonger durant tout le processus de l'ECBD. Suite à notre expérience de collaboration menée entre hydroécologues et informaticiens, nous conseillons d'ajouter deux étapes à l'ECBD en plus de l'étape préliminaire de conception. L'étape de filtration et de hiérarchisation des résultats issus de la fouille ne doit pas être négligée sous peine de laisser le thématicien submergé par d'innombrables résultats, en particulier si l'informaticien a fini son

contrat ! Cette étape se fait à l'aide de mesures d'intérêts à associer aux résultats (Geng et Hamilton, 2006). Le choix de ces mesures et de leurs éventuelles adaptations doit être discuté entre thématiciens et informaticiens. Ce choix doit être fait en fonction de ce qu'attendent les thématiciens et de ce qu'il est possible et pertinent de développer par les informaticiens. L'étape de validation doit inclure la reconnaissance d'une partie des résultats comme des résultats attendus par le thématicien, ainsi qu'un processus permettant d'écarter les résultats faux (Gibert et al., 2018). Cette étape est indispensable pour consolider le processus, avant de regarder les résultats inattendus. Les trois étapes à ajouter à l'ECBD, le dialogue permanent entre thématiciens et informaticiens et la nécessité de reprendre le processus de manière itérative jusqu'à aboutir à une « connaissance » nouvelle est schématisée en Figure 60.

L'aspect itératif du processus est indispensable et permet avec le temps à chacun de s'écouter, de mieux se comprendre, de mieux connaitre les possibilités et les contraintes de chacun, et d'acquérir un langage commun qui est la clé de réussite d'un projet pluridisciplinaire en général (Pohl, 2005) et d'un projet de fouille de données environnementales en particulier (Gibert et al., 2018). Dans un autre domaine, De Marsily & Fustec (1995) ont montré que la réussite du programme pluridisciplinaire PIREN Seine, qui n'est plus à démontrer, reposait sur une collaboration inscrite dans la durée et le partage d'un langage commun. La durée des contrats de recherche ne favorisent pas une collaboration sur du temps long. D'après Paasche et Österblom (2019), les pressions sur les chercheurs, concurrence dans la recherche de crédits, recherche de la reconnaissance par la course en avant des publications sur des sujets novateurs, ne favorisent pas non plus le travail pluridisciplinaire. Pourtant, la résolution des questions environnementales complexes nécessite la recherche de solutions pluridisciplinaires (Legay, 1999; Reid et Mooney, 2016). L'enjeu de la collaboration entre informaticiens et thématiciens est de maintenir la collaboration au-delà du développement d'une méthode de fouille prometteuse pour le domaine.

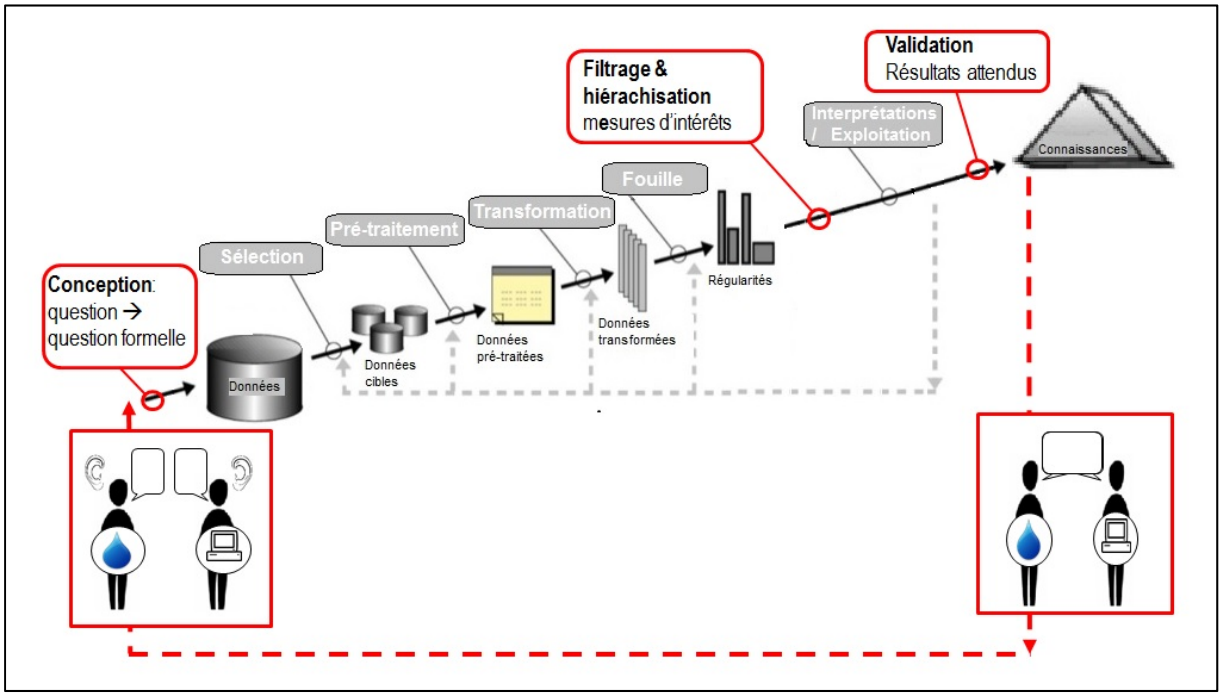


Figure 60 : Propositions d'ajout d'étapes – en rouge – à la schématisation de l'extraction de la connaissance des bases de données proposée par Fayyad et al. (1996)

L'application de méthodes de fouille non supervisée aux données rivières est possible, mais, afin d'obtenir des résultats probants, plusieurs prérequis, qu'il est parfois difficile de maitriser, sont nécessaires ainsi qu'un travail pluridisciplinaire approfondi et s'inscrivant dans la durée entre thématiciens, ici les hydroécologues, et informaticiens. Si ces conditions sont réunies alors le champ des possibles est ouvert.

1.3 Perspectives de la fouille non supervisée appliquée aux données sur les rivières

Nous avons proposé d'appliquer deux méthodes de fouille non supervisées aux deux questions suivantes, transformées en deux questions formelles qui ont permis de choisir des méthodes de fouille adaptées (chapitre II, 2.6).

Question 1 : existe-t-il des relations temporelles entre les résultats physico-chimiques et les indices biologiques en un même lieu? La question formelle retenue par les informaticiens est la recherche de motifs dans des séquences de données temporelles. La méthode de fouille choisie est l'extraction de motifs fermés partiellement ordonnés. Elle permet d'explorer des trajectoires temporelles. C'est la

méthode que nous avons approfondie, en l'étendant à un plus grand jeu de données, et dont nous développerons les apports, limites et perspectives dans les paragraphes suivants.

Question 2 : les variations des paramètres abiotiques d'une station de rivière peuvent-elles être reliées aux variations de fréquence d'expression de modalités de traits biologiques et écologqiues des taxons qui s'y trouvent ? La question formelle retenue par les informaticiens est la recherche de règles d'implication. La méthode de fouille choisie est l'analyse relationnelle de concepts Elle permet de prendre en compte les relations multiples entre les données abiotiques, les taxons et leurs traits. Nous avons appliqué cette méthode aux macro-invertébrés. Elle apparait donc comme adaptée à la question posée, mais pour changer d'échelle, nous avons encore deux problèmes à résoudre. Le premier est la mise en relation des stations et des listes taxonomiques disponibles à l'échelle nationale : ce lien n'est pas clairement établi dans la base de données. Le deuxième est la gestion de l'explosion du nombre de règles obtenues si la méthode est appliquée à un territoire plus grand, et à un plus grand nombre de taxons : poissons, diatomées, macrophytes à ajouter aux invertébrés.

Nica (2017) propose d'associer les deux méthodes et d'utiliser l'analyse relationnelle de concepts pour hiérarchiser les motifs fermés partiellement ordonnés. Ainsi dans un exemple issu de sa thèse (Figure 61), la méthode permet d'obtenir un treillis de tous les motifs qui ont précédé un indice IBGN en très bon état, représenté au sommet du treillis. Il y a eu des motifs avec un item de très bonne qualité et un motif avec un item de mauvaise qualité – niveau -1 sous le sommet. Parmi les motifs avec un item de très bonne qualité, il y a un motif composé d'un item matières organiques de très bonne qualité (MOOX en bleu) et un motif composé de deux items : un de très bonne qualité et un autre de bonne qualité – niveau -2 sous le sommet du treillis – ; etc... Pour l'instant cette méthode est restée expérimentale : elle a été appliquée aux données nationales, mais validée que sur un petit jeu de données : l'Alsace.

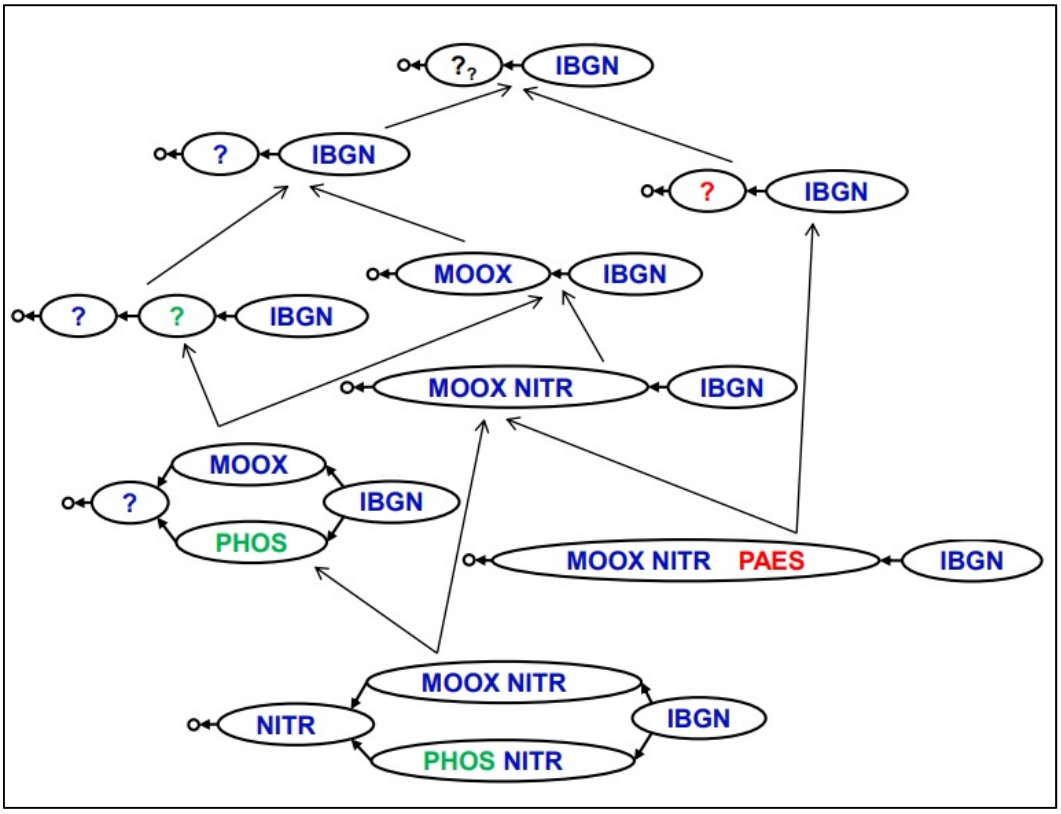


Figure 61 : exemple de treillis obtenu par une analyse relationnelle de concepts appliquée à des motifs temporels fermés partiellement ordonnés, sur le jeu de données de l'Alsace (Nica, 2017)

Nous avons utilisé les extractions de motifs fermés partiellement ordonnés sur des données séquentielles ordonnées dans le temps, mais cette méthode peut s'employer sur des données séquentielles ordonnées suivant d'autres grandeurs. Par exemple, la recherche de motifs spatiaux pourrait permettre d'explorer les relations amont – aval entre pressions liées à l'occupation du sol, et état physico-chimiques et biologiques. C'était une des questions supplémentaires que nous avions envisagées de soumettre à la fouille de données (chap II, appendix 1), mais que nous n'avons pas explorée. D'après Gibert et al. (2018), les méthodes de fouille peuvent s'appliquer à de multiples questions environnementales.

2 Le programme de fouille proposé PRESTOR : apports, limites et perspectives

2.1 PRESTOR un programme opérationnel et spécifiquement adapté aux données de surveillance des rivières

Pour pouvoir extraire des motifs nous avons développé à la fois une base de données relationnelle et le programme PRESTOR, s'appuyant sur la méthode de fouille élaborée par Fabrègue et al. (2014), que nous avons adaptée. La base de données accueille principalement les données de surveillance des rivières, de 2007 à 2013, mises à disposition par l'Agence Française de Biodiversité (AFB) ; elle a une extension géographique (logiciel libre PostgreSQL et module géographique PostGIS) et respecte le format SANDRE, dont sont issues notamment les tables taxons et paramètres. Elle inclut également les grilles SEQ-eau (MEDD et AE, 2003) et DCE (MEEM, 2012) permettant d'agréger les paramètres physico-chimiques en altérations et de discrétiser leurs valeurs en classes de qualité ou d'état. Par contre, elle n'inclut pas les seuils des états biologiques par indice biologique. L'état biologique est une information renseignée par station et par date, qui a été définie en fonction des seuils de 2012 (MEEM, 2012).

Par ailleurs PRESTOR propose à l'opérateur plusieurs critères pour choisir :

- les données d'entrée avant extraction : la période à prendre en compte (la totalité de la période disponible 2007-2013 ou une partie), la zone géographique (la France entière ou une seule HER), les grilles d'agrégations et de discrétisations des paramètres physico-chimiques (le SEQ-eau ou les grilles DCE), les altérations à prendre (toutes ou une partie) ;
- les caractéristiques des motifs à extraire en fixant la fréquence minimale d'extraction et la longueur de séquence à prendre.

PRESTOR fournit à l'opérateur non seulement les motifs extraits, mais également plusieurs fichiers associés de résultats permettant leur traitement. Nous avons décrit en chapitre III (2.4.4), ces fichiers résultats associés. En plus du rappel des critères ci-dessus choisis, ces fichiers fournissent le nombre de stations et de séquences disponibles par contexte (un indice biologique dans un état donné), les fréquences et nombre de stations, par contextes, vérifiant chaque motif. Nous avons enrichi ces résultats d'un fichier comptabilisant, par motif, le nombre global d'items et d'itemsets, le nombre de différents types d'items (exemple donné en Tableau 20). C'est un progrès, qui a profité aux deux derniers chapitres de ce travail, par rapport à une comptabilité manuelle de ces informations sur des motifs, qui étaient à l'origine seulement au format image.

Tableau 20 : Extrait d'un exemple de fichier de résultats comptabilisant les items, les itemsets et le type d'items par motif (extraction réalisée sur toute la France, avec les grilles DCE, pour une longueur de séquences 3 mois, et une fréquence minimale 0,48)

N° motif	items	item sets	BILO2=Bleu	BILO2=Vert	NUTRI=Bleu	POSPE=Bleu	SDP=Rouge
1	1	1	0	1	0	0	0
2	1	1	0	0	0	1	0
3	1	1	1	0	0	0	0
….							
226	2	2	0	0	0	1	1
227	2	1	0	0	0	1	1
228	4	3	0	0	0	1	3

Nous évoquions ci-dessus (1.1) la nécessité et parfois la difficulté de mettre en relation les données dans une base de données : l'état biologique en est un exemple. En effet, cet état est fonction à la fois de l'HER et du type de la masse d'eau. Pour mettre à jour cet état biologique sur la base des grilles DCE les plus récentes (MEEM, 2016), il faut renseigner les nouveaux seuils dans la base de données, ce qui est facilement réalisable. Mais il faut également que le type et l'HER de chaque station de surveillance soient renseignés, ce qui n'est pas le cas dans les données dont nous disposons.

PRESTOR est configuré pour effectuer des requêtes de séquences de données dont il a besoin directement dans la base. Ainsi une mise à jour de la base profitera aussi à l'extraction des motifs. Cet aspect peut paraître une évidence, mais il est plus facile, pour un informaticien, d'exporter d'abord les données nécessaires dans un fichier de type tableur, puis d'y appliquer la méthode de fouille choisie. Par contre, un thématicien pourra difficilement faire une nouvelle exportation de données, si celles –ci ont évoluées. Cet aspect est un des points à ne pas oublier de discuter entre informaticiens et thématiciens si l'objectif est de développer un outil durable.

L'ensemble des caractéristiques, décrites ci-dessus, permet à un non informaticien d'utiliser PRESTOR en autonomie, d'en extraire des motifs et d'avoir les résultats associés aidant à les interpréter. Dans notre cas, ces résultats incluent deux des mesures d'intérêts (la fréquence F et l'émergence E) et permettent de calculer les deux autres (la complexité C et la singularité S), sur lesquelles se basent la sélection des motifs les plus pertinents. A terme, nous souhaitons intégrer le calcul des mesures d'intérêts et la sélection des motifs dans PRESTOR.

2.2. Une méthode qualitative traitant des données discrétisées: critique des grilles de seuils utilisées

La méthode d'extraction des motifs est une méthode qualitative, aussi les données d'entrée doivent-elles être discrétisées. Nous avons choisi d'utiliser les classes de qualité et d'état. Pour limiter le nombre de paramètres physico-chimiques à considérer, nous les avons agrégés en altérations du SEQ-eau ou des grilles DCE.

Pour la discrétisation des indices biologiques, comme expliqué ci-dessus, nous utilisons des seuils des grilles DCE de 2012. Ces seuils ont peu évolués depuis excepté pour l'indice I2M2, dont l'utilisation sur les réseaux de surveillance est plus récente (2007) et dont la dernière décennie a été consacrée au calage de ses seuils par grand bassin (*comm. perso. S. Dembski, AFB, 16 janvier 2019*). Ceci pourrait expliquer pourquoi nous avons trouvé pour cet indice de motifs réellement caractéristiques, contrairement aux motifs pour l'IBGN, qui est basé sur le même groupe faunistique. Il faudrait donc changer la discrétisation de l'état des résultats de l'I2M2. En général, quel que soit l'indice biologique, les motifs les plus discriminants

sont obtenus pour les classes de qualité extrêmes, très bonnes et mauvaises, ce qui met en évidence la difficulté de fixer des valeurs-seuils de classe pertinente, difficulté soulignée par Birk et al. (2012). Ont fait exception à cette règle les motifs extraits pour l'IBMR en état médiocre avec le SEQ-eau et ceux pour l'IBD de bonne qualité avec les grilles DCE.

Concernat la comparaison entre les deux grilles de seuils utilisées (SEQ-eau et grilles DCE), nous avons obtenu des motifs moins intéressants avec les grilles DCE car ils étaient moins complexes, moins singuliers, comptaient moins de motifs dominants, que ceux abtenus avec le SEQ-eau. Cette différence s'explique par le nombre d'altérations et le nombre de paramètres moins importants dans les grilles DCE que dans le SEQ-eau : 6 altérations contre 14, 58 paramètres contre 183. Entre les grilles DCE de 2012 et celles les plus récentes de 2016, déjà évoquées ci-dessus, le nombre de paramètres a légèrement augmenté : le nombre de substances prioritaires et dangereuses prioritaires (SDP) est passés de 38 à 45 ; le nombre de polluants spécifiques de l'état écologique (POSPE) est passé de 9 à 15 ou 20, suivant les bassins des Agences de l'Eau. Ainsi le nombre global de paramètres est passé de 65 à 78 ou 83 suivant le bassin, ce qui reste bien inférieur à 183 pour le SEQ-eau. Le SEQ-eau différencie les principaux micropolluants en cinq altérations basées sur les types de polluants – micropolluants minéraux, organiques hors pesticides, HAP, PCB et pesticides – au lieu de deux – SDP et POSPE – et les nutriments en trois altérations –nitrates, matières azotés hors nitrates et matières phosphorées- au lieu d'une seule. Cette différentiation plus poussée du SEQ-eau explique que nous obtenions des motifs plus complexes et elle apparait plus adapté à l'objectif d'identification des pressions mis en cause dans la non atteinte du bon état écologique. La limitation du nombre de paramètres pris en compte dans les altérations correspondant à la pollution en matières organiques –altération MOOX pour matières organiques et oxydables du SEQ-eau, et BILO2 pour bilan en oxygène des grilles DCE – peut expliquer les écarts entre ces deux altérations dans les motifs. Dans les grilles DCE, la demande chimique en oxygène (DCO) et l'azote réduit (ammonium et azote Kjeldahl) ne sont plus pris en compte dans cette altération. Et, en effet, dans les motifs extraits pour les différentes longueurs de séquences, si ce type d'altérations est régulièrement trouvé en qualité moyenne avec le SEQ-eau, elle

ne l'est quasiment pas avec les grilles DCE. Les seuils de deux paramètres ont également évolués de façon significative entre le SEQ-eau et les grilles DCE. Le seuil du bon état pour les nitrates est passé de 10 mg/L à 50 mg/. Pour le glyphosate, herbicide très utilisé (environ 40 % du volume mondial des pesticides (EPA, 2011)) et qui fait souvent la une de l'actualité depuis plusieurs années, le seuil du bon état est de 0,4 µg/L dans le SEQ-eau, puis cette molécule est absente des grilles DCE de 2009 et 2012, pour réapparaitre dans les grilles de 2016 à 28 µg/L. Sa toxicité sur des espèces non ciblées notamment aquatiques (*Daphnia magna*) a été démontré (Sihtmäe et al., 2013). Nous n'avons pas obtenu de motifs avec des nutriments autres que très bons et bons, bien que ces autres états aient été présents dans les données dont nous disposions. Chaque citoyen peut s'interroger sur l'hypocrisie d'afficher une politique ambitieuse : restaurer le bon état écologique et, en même temps, augmenter les seuils des deux paramètres dont les pollutions sont difficiles à maitriser.

Il existe peu de travaux comparant les seuils utilisés par les différents pays Européens. De plus, la comparaison n'est pas aisée car les pays n'utilisent pas les mêmes modes d'agrégations : certains prennent les moyennes ou les médianes annuelles des valeurs mesurées, alors que la France retient les percentiles 90 (ou 10 dans certains cas). Pourtant les seuils français du bon état, des grilles DCE, ne sont pas les valeurs parmi les plus élevées : 0,5 mg/L pour l'ammonium alors que l'ensemble des seuils européens appartiennent à l'intervalle [0,05 ; 1,6 mg/L] (Claussen et al., 2012), 0,5 mg/L pour les phosphates comparé à l'intervalle européen de [0,05 ; 1 mg/L] (Arle et al., 2016), excepté pour l'azote global : 12 mg/L (incluant ammonium, nitrites et nitrates) comparé à l'intervalle européen de [0,7 ; 12mg/L] (Claussen et al., 2012). Par contre, les seuils actuels utilisés pour les substances prioritaires (SDP) sont partagés par tous les pays européens.

Concernant les altérations des pesticides du SEQ-eau et celles des substances prioritaires des grilles DCE, nous n'avons pas obtenu de motif discriminant des états biologiques, mis à part pour les pesticides pour l'IBMR en mauvais état, pour des extractions faites pour 24 mois. Les pesticides sont reconnus comme largement répandus et représentant une pression majeure sur les

écosystèmes aquatiques (Millenium Ecosystem Assessment Programm, 2005). Mais il est à noter qu'en comparant les quantités analysées de ces micropolluants et leurs seuils de classes (Honda, 2019) sur le quart nord est français (représentant 17% des stations nationales), nous nous sommes aperçus que certaines limites de quantifications pouvaient être égales voire supérieures à des seuils déclassant. Ainsi pour l'insecticide Chlorpyrifos-éthyl, la limite SEQ-eau entre la classe moyenne et la classe médiocre est de 0,005 µg/L. Or 47 % des valeurs disponibles ont cette même valeur qui est la limite de quantification. Ainsi les 47 % des valeurs de cet insecticide, qui ont été mesurées avec cette valeur de quantification, seront classés en qualité médiocre tout comme l'altération pesticides, qui sera à minima la plus mauvaise qualité de l'ensemble des paramètres mesurés. Il en est de même pour le fongicide Chlorothalonil, dont le seuil entre la classe moyenne et la classe médiocre est de 0,04 µg/L et dont 77 % des valeurs disponibles ont cette même valeur qui est en fait la limite de quantification.

Pour s'affranchir des limites des seuils dont nous disposons, il faudrait utiliser les données brutes, mais alors la méthode d'extraction de motifs temporels n'est plus adaptée. Nous avons commencé à tester une nouvelle méthode : l'analyse interactive de clustering sous contraintes capable de traiter des séquences temporelles de données quantitatives (Braud et al., 2019, en soumission).

3 Les motifs extraits aux échelles nationale et d'une HER : apports, limites et perspectives

C'est la première fois, à notre connaissance, qu'est proposée l'extraction de successions temporelles d'altérations partagées par des ensembles de stations de rivières surveillées à différentes dates. Après avoir abordé cet aspect novateur, nous ferons la synthèse des apports, limites et perspectives des motifs temporels extraits à l'échelle nationale et d'une HER.

3.1 Une première : des successions temporelles d'altérations précédant un état biologique

Les résultats produits par PRESTOR sont stables, et exploitables par des non informaticiens, deux des conditions importantes à vérifier lorsqu'on applique une méthode de fouille (Gibert et al., 2018).

Les motifs sont la synthèse des successions temporelles d'altérations physico-chimiques les plus fréquentes, partagée par un ensemble de séquences, ici des stations de surveillance à différentes dates, qui ont précédé un état biologique donné. Ils ont le double avantage de représenter à la fois la chronologie des événements et l'association de ces événements.

Ils sont facilement compréhensibles, sous réserve de ne pas être trop grands. Les plus grands motifs présentés dans ce travail sont le motif n° 165, qui compte 139 items et le motif n° 171, qui compte 245 items (ils ont été extraits sur l'HER Alsace, pour des séquences de 60 mois, en se limitant aux altérations des macro-polluants ; le motif n° 165 est en annexe 9), mais c'est une exception.

Mais, bien sûr, la méthode a ses limites. Bien que performant, le programme utilisé ne peut extraire des motifs que dans des conditions bornées. A l'échelle nationale, nous avons dû procéder par éliminations successives des altérations dominantes dans les premières extractions, pour obtenir des motifs avec d'autres items que la majorité des macro-polluants en bon état. Ce type de pré-traitements des données est courant en fouille de données (Gibert et al., 2008) tout comme en bio-statistiques (Serrano Balderas et al., 2017). La plus grande longueur de séquences a été de 24 mois, pour laquelle nous n'avons pas pu descendre en dessous d'une fréquence minimale de 0,6. En fouille de données, le nombre de résultats produits peut largement dépasser le nombre d'objets à décrire : les informaticiens parlent d'explosion combinatoire de l'information à traiter. La pré-sélection des données et la limitation des critères d'extraction ont été nos solutions pour rester dans les limites de notre programme. D'autres moyens peuvent être envisagés pour réduire les coûts en temps et en espace mémoire nécessaire au

programme de fouille tels qu'ajouter des contraintes de sélection sur les données d'entrée (Geng et Hamilton, 2006). Nous pourrions par exemple imaginer d'extraire les motifs qui n'ont que des altérations médiocres et mauvaises pour les macro-polluants. En effet, il s'agit d'items que nous n'avons pas trouvés dans les motifs extraits, exceptés pour les particules en suspension. Mais cela nécessiterait d'autres développements dans PRESTOR.

Un des biais de la méthode est également d'extraire plus de motifs pour les petits jeux de données. Ce problème a été amplifié par le jeu de données utilisé : il présente une répartition déséquilibrée des données, notamment peu de stations en mauvais états pour les indices IBGN et IBMR. Mais ce jeu de données a l'avantage d'être national.

3.2 Apports, limites et perspectives des motifs temporels extraits à l'échelle nationale

Les motifs extraits à l'échelle nationale nous ont permis d'apporter des réponses aux questions posées dans l'introduction.

Les successions de pressions temporelles sur plusieurs mois à années permettent-elles d'expliquer les réponses des différents indices biologiques à un instant donné?

Dans les motifs obtenus sur toute la France, les altérations physico-chimiques sont majoritairement en meilleur état que les indices biologiques qu'ils précédent que ce soit avec le SEQ-eau ou avec les grilles DCE, et quelles que soient les longueurs de séquences. Deux altérations du SEQ-eau font exception : les matières en suspension, trouvées dans des états équivalents et les pesticides, trouvés en qualité médiocre ou mauvaise pour tous les états biologiques de très bon à mauvais. C'est aussi le cas des substances prioritaires des grilles DCE, trouvées en mauvais état pour tous les états biologiques. Les états biologiques extrêmes : très bons et mauvais, et médiocre pour l'IBMR, sont bien caractérisés par des motifs émergents, singuliers, complexes au contraire des classes intermédiaires pour lesquelles les motifs trouvés sont le plus souvent non émergents, peu complexes, peu singuliers,

voire triviaux (par exemple un seul item de nitrates en bon état en 24 mois qui précède un IBD moyen).

Certaines de ces successions peuvent-elles être identifiées comme caractéristiques de réponses types et donc servir à prédire une réponse biologique?

Les indices biologiques en très bon état étaient précédés de motifs composés majoritairement d'altérations dans les classes très bonne et bonne et se maintenant dans le temps. C'est le cas principalement pour les matières phosphorées et, dans une moindre mesure, pour les matières organiques, ainsi que de l'acidité. A l'opposé, pour les mauvais états biologiques, il existe deux types de motifs :

- des motifs avec des altérations dégradées multiples et chroniques composés à la fois de macro-polluants et de micropolluants, se répétant plusieurs fois;
- ou une seule apparition d'une altération dégradée en macro-polluants, parmi d'autres altérations non dégradées.

Les altérations en macro-polluants retrouvées dans les motifs extraits pour de mauvais états biologiques sont principalement celles des nutriments et des matières organiques, comme chez d'autres auteurs (Comte et al., 2010; Lemm et Feld, 2017; Stendera et al., 2012; Villeneuve et al., 2015). La réponse des communautés macrophytiques à la pollution organique a déjà été démontrée (Haury et al., 2006; Thiébaut, 2006).

Existe-t-il des différences entre motifs extraits par indice biologique en fonction des longueurs de séquences considérées ?

Pour les altérations liées aux micropolluants, connus pour avoir des effets toxiques rapides sur les taxons aquatiques, nous observons, pour tous les indices biologiques, dans les motifs extraits dès 3 mois, l'apparition des items de l'altération pesticides de qualité médiocre et mauvaise du SEQ-eau ou de l'altération des substances prioritaires en mauvais état des grilles DCE. Mais, ces items sont peu discriminants des états biologiques, mis à part pour l'IBMR et l'IPR en mauvais état

pour la longueur de séquences 24 mois. C'est seulement pour l'IBGN en mauvais état que les motifs, dès 3 mois et jusqu'à 24 mois, comportent les items des micropolluants organiques hors pesticides de qualité moyenne et des micropolluants minéraux de qualité médiocre.

Pour les altérations des macro-polluants liés aux matières organiques, nitrates et autres matières azotées, matières phosphorées et particules en suspension, nous avons observé des différences dans les motifs extraits à différentes longueurs de séquences pour les indices IBD, IBGN et IBMR. Dans l'ensemble des motifs de l'IBMR et de l'IBGN, ces altérations sont absentes ou en très faible nombre pour les extractions aux longueurs de séquences courtes, 3 et 6 mois, et à l'inverse importantes au-delà de 12 mois. Alors que ces altérations sont présentes dans les motifs extraits pour l'IBD dès 3 mois.

Nous n'avons pas observé de différences significatives en fonction des longueurs de séquences dans les motifs extraits pour les indices IPR et I2M2. Pour les poissons, extraire des motifs limités aux seules altérations physico-chimiques ne semble pas suffisant et il faudrait y intégrer les pressions hyromorphologiques auxquelles ils sont reconnus comme sensibles (Dahm et al., 2013; Marzin et al., 2012; Reyjol et al., 2008; Villeneuve et al., 2015). Nous avons déjà discuté ci-dessus (chapitre IV, 4) du cas de l'I2M2 pour lequel nous ne disposions pas d'états stabilisés.

Mais l'extraction de motifs à grande échelle sur le territoire français métropolitain limite leur pertinence et les interprétations qui peuvent en être faites, ainsi que la longueur maximale des séquences possibles à 24 mois. Aussi proposons-nous d'appliquer cette méthode à l'échelle des hydro-éco-régions (HER).

3.3 Apports, limites et perspectives des extractions de motifs à l'échelle d'une HER

Nous voulions tester si travailler à l'échelle d'une HER nous permettait d'allonger les longueurs de séquences. En effet, en travaillant à l'échelle nationale nous avons atteint les limites d'extractions de PRESTOR à 24 mois. Au-delà les tentatives d'extraction échouaient. Sous réserve de restreindre les données d'entrée

à un seul indices biologique, ici l'I2M2, et à une partie des altérations, ici les macro-polluants, et de ne prendre qu'une HER, ici l'Alsace, l'extraction de motifs a abouti pour une durée de 60 mois ; ce qui correspond à l'ensemble des données dont nous disposions.

Nous voulions également affiner notre méthode de sélection des motifs et les attribuer à des groupes de stations. Pour cela, outre l'utilisation des quatre mesures d'intérêts et de leur combinaison, nous avons appliqué une classification des motifs sur la base des séquences (stations-dates) qu'ils partagent. Ainsi nous avons pu proposer des groupes de motifs, caractérisant des stations et correspondant à l'évolution ou la stabilité de l'état biologique de l'I2M2.

La démarche reste à affiner : il serait plus pertinent de faire une classification des stations sur la base de leurs motifs communs. Pour l'instant, le programme PRESTOR d'extraction des motifs ne permet pas de sélectionner les séquences sur la base d'une évolution d'un état biologique, que ce soit sa dégradation ou son amélioration. A terme, c'est ce que nous souhaitons faire. Ainsi, alors que l'outil diagnostic proposé par Mondy et Usseglio-Polatera (2013) sur les invertébrés ou Larras et al (2017) sur les diatomées permet d'identifier les pressions à l'échelle d'une station, nous pourrions proposer une méthode qui permette d'identifier des groupes de stations ayant les mêmes types de réponse biologiques, et donc le même potentiel de résilience, à des pressions similaires. Pour cela, il faudrait ajouter aux données dont nous disposons les données des pressions morphologiques et hydrologiques, d'autant plus importantes que ces paramètres sont déterminants dans le contexte actuel de changement climatique.

CONCLUSION

Les données de surveillance de l'état des rivières sont, à présent, massives et complexes. Nous avons montré que les méthodes de fouille non supervisées peuvent s'appliquer à ces données, sous réserve qu'elles soient structurées dans une base de données relationnelle et que soit mis en place un processus plus large d'extraction de connaissances, basé sur une collaboration rapprochée, à la fois interactive et itérative, entre hydroécologues et informaticiens. Dans ces conditions, il est possible d'obtenir des résultats pertinents, aisément compréhensibles et pouvant aider à la gestion des masses d'eau.

En particulier, nous avons développé le programme PRESTOR, qui permet l'extraction de motifs temporels partiellement ordonnés à partir des séquences d'altérations physico-chimiques qui précèdent un état biologique, dans l'objectif d'identifier les pressions qui ont conduit à cet état. L'identification des pressions, d'autant plus difficile dans un contexte multi-perturbé, est une étape clé de la restauration du bon état écologique qui est l'objectif visé par la Directive Cadre européenne sur l'Eau (2000).

PRESTOR a été spécifiquement adapté aux données rivières et est utilisable par un non informaticien. Il est basé sur une méthode de fouille qualitative qui nécessite de discrétiser les données en entrée : nous utilisons les classes d'état biologiques (seuils de 2012) et nous avons testé les grilles du Système d'Evaluation de la Qualité de l'eau (SEQ-eau, 2003) et de la DCE (2012) pour les données physico-chimiques. A l'échelle de la France métropolitaine, sur la période 2007-2013, nous avons extrait les motifs pour les états des cinq indices biologiques basés sur les quatre groupes invertébrés (IBGN et I2M2), poissons (IPR), diatomées (IBD) et macrophytes (IBMR).

Nous avons obtenu des motifs discriminants pour les classes extrêmes des très bon et mauvais états écologiques, mais rarement pour les états écologiques intermédiaires bon, moyen et médiocre, ce qui souligne la difficulté de définir des valeurs de seuils pertinentes, en particulier pour le bon état. Les motifs extraits pour

les très bons états sont caractérisés par des résultats physico-chimiques stables dans le temps en très bon ou bon états, excepté pour les pesticides et substances prioritaires et dangereuses prioritaires, qui ont été trouvées en état dégradé quels que soient les états biologiques. Les motifs extraits pour les mauvais états biologiques sont de deux types : soit des motifs caractérisant une situation perturbée chroniquement en cumulant des altérations dégradées à la fois pour les macro- et micropolluants ; soit des motifs caractérisant une situation globalement bonne mais avec une seule altération dégradée. Nous avons testé l'extraction de motifs pour différentes longueurs de séquences de 3 à 24 mois et constaté des différences à la fois dans les altérations qui les constituent et les indices biologiques pour lesquels ils sont extraits. La majorité des altérations des micropolluants sont présentes dans les motifs extraits dès 3 mois. Les altérations des macro-polluants sont présentes dans les motifs extraits pour l'indice IBD dès 3 mois, alors qu'ils apparaissent dans les motifs extraits pour l'IBGN et l'IBMR à partir de 12 mois. Les altérations des micropolluants minéraux et organiques hors pesticides ont essentiellement été trouvées dans les motifs extraits pour l'IBGN en mauvais état. Il est possible que les résultats moins probants obtenus pour l'I2M2 soient dus aux seuils utilisés pour discrétiser l'état biologique de cet indice, non encore stabilisés en 2012 et pour l'IPR, à l'absence de prise en compte des altérations hydromorphologiques dans nos données. Le SEQ-eau apparait comme plus pertinent pour discrétiser les données physico-chimiques, en particulier, parce qu'il propose plus d'altérations et permet donc de les distinguer.

L'extraction de motifs à grande échelle sur le territoire français métropolitain limite leur pertinence et les interprétations qui peuvent en être faites, notamment pour travailler à l'échelle des stations. De plus, la méthode atteint ses limites : les fréquences minimales sont forcément hautes. Aussi proposons-nous d'appliquer cette méthode à l'échelle des hydro-éco-régions pour catégoriser des stations souffrant des mêmes pressions et dans des états biologiques similaires.

C'est la première fois que sont proposés des motifs représentant la succession temporelle d'altérations physico-chimiques qui ont précédé un état biologique. Ils sont facilement exploitables. A terme, les pistes d'amélioration de notre programme spécifique PRESTOR sont d'intégrer les données des altérations

morphologiques et hydrologiques – ces dernières étant d'autant plus importantes dans le contexte actuel de changement climatique – et d'extraire des motifs non pas pour un indice biologique dans un état donné, mais pour un indice biologique ayant connu une amélioration ou à l'inverse une dégradation.

BIBLIOGRAPHIE

Adriaenssens, V., Goethals, P.L.M., Charles, J., De Pauw, N., 2004. Application of Bayesian Belief Networks for the prediction of macroinvertebrate taxa in rivers. Ann. Limnol. 40, 181–191. https://doi.org/10.1051/limn/2004016

AFNOR, 2016. Qualité de l'eau: prélèvements des macro-invertébrés aquatiques en rivières peu profondes. Norme Française NF T 90-333.

AFNOR, 2010. Qualité de l'eau: traitement au laboratoire d'échantillons contenant des macro-invertébrés de cours d'eau. Norme Française Expérimentale XP T 90-388.

AFNOR, 2007. Qualité de l'eau: détermination de l'Indice Biologique Diatomées (IBD). Norme Française NF T90-354.

AFNOR, 2004a. Qualité de l'eau: détermination de l'Indice Biologique Global Normalisé (IBGN); Norme Française NF T90-350.

AFNOR, 2004b. Qualité de l'eau: détermination de l'Indice poissons rivière (IPR). Norme Française NF T90-344.

AFNOR, 2003. Qualité de l'eau: détermination de l'Indice Biologique Macrophytique en Rivière (IBMR). Norme Française NF T90-395.

AFNOR, 2002. Qualité de l'eau : détermination de l'Indice Oligochètes de Bioindication des Sédiments (IOBS). NF T90-390.

Agrawal, R., Srikant, R., 1995. Mining sequential patterns, in: Proceedings of the Eleventh International Conference on Data Engineering. IEEE Comput. Soc. Press, pp. 3–14. https://doi.org/10.1109/ICDE.1995.380415

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules, in: Proc. 20th IVLDB Conference.

Allan, J.D., 2004a. Influence of land use and landscape setting on the ecological status of rivers. Limnetica 23, 187–198. https://doi.org/10.1146/annurev.ecolsys.35.120202.110122

Allan, J.D., 2004b. Influence of land use and landscape setting on the ecological status of rivers. Annu. Rev. Ecol. Evol. Syst. 35, 257–284. https://doi.org/10.1146/annurev.ecolsys.35.120202.110122

Altenburger, R., Ait-Aissa, S., Antczak, P., Backhaus, T., Barceló, D., Seiler, T.B., Brion, F., Busch, W., Chipman, K., de Alda, M.L., de Aragão Umbuzeiro, G., Escher, B.I., Falciani, F., Faust, M., Focks, A., Hilscherova, K., Hollender, J., Hollert, H., Jäger, F., Jahnke, A., Kortenkamp, A., Krauss, M., Lemkine, G.F., Munthe, J., Neumann, S., Schymanski, E.L., Scrimshaw, M., Segner, H., Slobodnik, J., Smedes, F., Kughathas, S., Teodorovic, I., Tindall, A.J., Tollefsen, K.E., Walz, K.H., Williams, T.D., Van den Brink, P.J., van Gils, J., Vrana, B., Zhang, X., Brack, W., 2015. Future water quality monitoring - Adapting tools to

deal with mixtures of pollutants in water resource management. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2014.12.057

Anselin, L., 1995. Local Indicators of Spatial Association-LISA. Geogr. Anal. 27, 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Archaimbault, V., Usseglio-Polatera, P., Garric, J., Wasson, J.-G., Babut, M., 2010. Assessing pollution of toxic sediment in streams using bio-ecological traits of benthic macroinvertebrates. Freshw. Biol. 55, 1430–1446. https://doi.org/10.1111/j.1365-2427.2009.02281.x

Arle, J., Claussen, U., Müller, P., 2011. Comparison of Environmental Quality Objectives, Threshold Values or Water Quality Targets Set for the Demands of EuropeanWater Framework Directive; Report for the CISWorking Group A "ECOSTAT."

Arle, J., Mohaupt, V., Kirst, I., 2016. Monitoring of Surface Waters in Germany under the Water Framework Directive — A Review of approaches, methods and results. Water 8, 217. https://doi.org/10.3390/w8060217

Arnaud, F., Piégay, H., Schmitt, L., Rollet, A.J., Ferrier, V., Béal, D., 2015. Historical geomorphic analysis (1932-2011) of a by-passed river reach in process-based restoration perspectives: The Old Rhine downstream of the Kembs diversion dam (France, Germany). Geomorphology 236, 163–177. https://doi.org/10.1016/j.geomorph.2015.02.009

Arts, G.H.P., Belgers, J.D.M., Hoekzema, C.H., Thissen, J.T.N.M., 2008. Sensitivity of submersed freshwater macrophytes and endpoints in laboratory toxicity tests. Environ. Pollut. 153, 199–206. https://doi.org/10.1016/j.envpol.2007.07.019

Aubry, P., Piegay, H., 2001. Spatial autocorrelation analysis in geomorphology: Definitions and tests. Geogr. Phys. Quat. 55, 111_129.

Baran, P., Lek, S., Delacoste, M., Belaud, A., 1996. Stochastic models that predict trout population density or biomass on a mesohabitat scale. Hydrobiologia 337, 1–9. https://doi.org/10.1007/BF00028502

Beketov, M.A., Foit, K., Schäfer, R.B., Schriever, C.A., Sacchi, A., Capri, E., Biggs, J., Wells, C., Liess, M., 2009. SPEAR indicates pesticide effects in streams - Comparative use of species- and family-level biomonitoring data. Environ. Pollut. https://doi.org/10.1016/j.envpol.2009.01.021

Benzécri, J.P., 1973. L'analyse des données. T.I : la taxonomie. T.n : l'Analyse des correspondances., Dunod. ed. Paris.

Berenzen, N., Schulz, R., Liess, M., 2001. Effects of chronic ammonium and nitrite contamination on the macroinvertebrate community in running water microcosms. Water Res. 35, 3478–3482. https://doi.org/10.1016/S0043-1354(01)00055-0

Berrahou, L., Lalande, N., Serrano Balderas, E.C., Molla, G., Berti-Équille, L., Bimonte, S., Bringay, S., Cernesson, F., Grac, C., Ienco, D., Le Ber, F., Teisseire, M., 2015. A quality-aware spatial data warehouse for querying hydroecological data. Comput. Geosci. 85, 126–135. https://doi.org/10.1016/j.cageo.2015.09.012

Bertaux, A., 2010. Treillis de Galois pour les contextes multi-valués flous. Application à l'etude des traits de vie en hydrobiologie. Thèse. Univesité de Strasbourg.

Berti-Equille, L., 2004. Un état de l'art sur la qualité des données. Ingénierie des systèmes d'information 9, 117–143. https://doi.org/10.3166/isi.9.5-6.117-143

Bimonte, S., Boulil, K., Braud, A., Bringay, S., Cernesson, F., Dolques, X., Fabrègue, M., Grac, C., Lalande, N., Le Ber, F., Teisseire, M., 2015. A decisional system for analysing water quality of watercourses. Ing. des Syst. d'Information 20. https://doi.org/10.3166/ISI.20.3.143-167

Birk, S., Bonne, W., Borja, A., Brucet, S., Courrat, A., Poikane, S., Solimini, A., Van De Bund, W., Zampoukas, N., Hering, D., 2012. Three hundred ways to assess Europe's surface waters: An almost complete overview of biological methods to implement the Water Framework Directive. Ecol. Indic. 18, 31–41. https://doi.org/10.1016/j.ecolind.2011.10.009

Blanck, H., 2002. Critical review of procedures and approaches used for assessing pollution- induced community tolerance (PICT) in biotic communities. Hum Ecol Risk Assess 8, 1003–10034.

Blard-Zakar, A., Michon, J., 2018. Bulletin n°3: rapportage 2016 des données au titre de la DCE.

Bonada, N., Prat, N., Resh, V.H., Statzner, B., 2006. Developments in aquatic insecte biomonitoring: A comparative analysis of recent approaches. Annu. Rev. Entomol. 51, 495–523. https://doi.org/10.1146/annurev.ento.51.110104.151124

Borja, A., Dauer, D.M., 2008. Assessing the environmental quality status in estuarine and coastal systems: Comparing methodologies and indices. Ecol. Indic. 8, 331–337. https://doi.org/10.1016/j.ecolind.2007.05.004

Bornette, G., Henry, C., Barrat, M.-H., Amoros, C., 1994. Theoretical habitat templets, species traits, and species richness: aquatic macrophytes in the Upper Rhône River and its floodplain. Freshw. Biol. 31, 487–505. https://doi.org/10.1111/j.1365-2427.1994.tb01753.x

Bouleau, G., 2007. La gestion française des rivières et ses indicateur à l'epreuve de la directive cadre. Thèse. AgroParisTech.

Boulil, K., Le Ber, F., Bimonte, S., Grac, C., Cernesson, F., 2014. Multidimensional modeling and analysis of large and complex watercourse data: An OLAP-based solution. Ecol. Inform. 24, 90–106. https://doi.org/10.1016/j.ecoinf.2014.07.001

Bouzeghoub, M., Mosseri, R., 2017. Les Big Data à découvert, CNRS éditi. ed.

Braud, A., Gançarski, P., Grac, C., Le Ber, F., 2019. Classification de séries temporelles hétérogènes pour le suivi de l'état des cours d'eau. Conférence Francoph. EGC - Extr. Gest. des Connaissances-, 27-31/01/2020, Bruxelles, B, en soumission.

Brogna, D., 2017. Forest cover impact on water related ecosystem services: methods and application at the regional scale (Wallonia, Belgium). Université de Namur.

Bruno, D., Belmar, O., Sánchez-Fernández, D., Guareschi, S., Millán, A., Velasco, J., 2014. Responses of Mediterranean aquatic and riparian communities to human pressures at different spatial scales. Ecol. Indic. 45, 456–464. https://doi.org/10.1016/j.ecolind.2014.04.051

Claussen, U., Müller, P., Arle, J., 2012. WFD CIS ECOSTAT WG A Report " Comparison of envi- ronmental quality Objectives , threshold values or water quality targets Set for the demands of the European Water Framework Directive ". Version 1 . Internal report , 2012 .

Combroux, I., Bornette, G., Willby, N.J., Amoros, C., 2001. Regenerative strategies of aquatic plants in disturbed habitats: The role of the propagule bank. Arch. fur Hydrobiol. 152, 215–235.

Comte, L., Lek, S., de Deckere, E., de Zwart, D., Gevrey, M., 2010. Assessment of stream biological responses under multiple-stress conditions. Environ. Sci. Pollut. Res. 17, 1469–1478. https://doi.org/10.1007/s11356-010-0333-z

D'heygere, T., Goethals, P.L.M., De Pauw, N., 2006. Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks. Ecol. Modell. 195, 20–29. https://doi.org/10.1016/j.ecolmodel.2005.11.005

D'heygere, T., Goethals, P.L.M., De Pauw, N., 2003. Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. Ecol. Modell. 160, 291–300. https://doi.org/Pii S0304-3800(02)00260-0

Dahm, V., Hering, D., Nemitz, D., Graf, W., Schmidt-Kloiber, A., Leitner, P., Melcher, A., Feld, C.K., 2013. Effects of physico-chemistry, land use and hydromorphology on three riverine organism groups: A comparative analysis with monitoring data from Germany and Austria. Hydrobiologia 704, 389–415. https://doi.org/10.1007/s10750-012-1431-3

Dakou, E., Goethals, P.L.M., D'heygere, T., Dedecker, A.P., Gabriels, W., De Pauw, N., Lazaridou-Dimitriadou, M., 2006. Development of artificial neural network models predicting macroinvertebrate taxa in the river Axios (Northern Greece). Ann. Limnol. J. Limnol. 42, 241–250. https://doi.org/Doi 10.1051/Limn/2006025

De Lange, H.J., Sala, S., Vighi, M., Faber, J.H., 2010. Ecological vulnerability in risk assessment — A review and perspectives. Sci. Total Environ. 408, 3871–3879. https://doi.org/10.1016/j.scitotenv.2009.11.009

De Marsily, G., Fustec, E., 1995. Le programme CNRS Piren-Seine : une action de recherche pluridisciplinaire et mulitpartenariale sur le fonctionnement global d'un bassin fluvial.

Demars, B.O.L., Potts, J.M., Trémolières, M., Thiébaut, G., Goucelin, N., Nordmann, V., 2012. River macrophyte indices: Not the Holy Grail! Freshw. Biol. 57, 1745–1759. https://doi.org/10.1111/j.1365-2427.2012.02834.x

Dolédec, S., Phillips, N., Scarsbrook, M., Riley, R.H., Townsend, C.R., 2006. Comparison of structural and functional approaches to determining landuse effects on grassland stream invertebrate communities. J. North Am. Benthol. Soc. 25, 44–60. https://doi.org/10.1899/0887-3593(2006)25[44:COSAFA]2.0.CO;2

Dolédec, S., Statzner, B., 2008. Invertebrate traits for the biomonitoring of large European rivers: an assessment of specific types of human impact. Freshw. Biol. 53, 617–634. https://doi.org/10.1111/j.1365-2427.2007.01924.x

Dolques, X., Le Ber, F., Huchard, M., Grac, C., 2016. Performance-friendly rule extraction in large water data-sets with AOC posets and relational concept analysis. Int. J. Gen. Syst. 45, 187–210. https://doi.org/10.1080/03081079.2015.1072927

Dray, S., Dufour, A.-B., 2007. The ade4 Package: Implementing the Duality Diagram for Ecologists. J. Stat. Softw. 22, 18–36.

Dunham, M.H., 2003. Data Mining. Introductory and Advanced Topics, Prentice H. ed.

Durance, I., Ormerod, S.J., 2007. Climate change effects on upland stream macroinvertebrates over a 25-year period. Glob. Chang. Biol. 13, 942–957. https://doi.org/10.1111/j.1365-2486.2007.01340.x

Džeroski, S., 2001. Applications of symbolic machine learning to ecological modelling. Ecol. Modell. 146, 263–273. https://doi.org/10.1016/S0304-3800(01)00312-X

Džeroski, S., Demšar, D., Grbović, J., 2000. Predicting chemical parameters of river water quality from bioindicator data. Appl. Intell. 13, 7–17. https://doi.org/10.1023/A:1008323212047

Džeroski, S., Grbović, J., Walley, W.J., Kompare, B., 1997. Using machine learning techniques in the construction of models. II. Data analysis with rule induction. Ecol. Modell. 95, 95–111. https://doi.org/10.1016/S0304-3800(96)00029-4

Elbrächter, M., 1977. On population dynamics in multi-species cultures of diatoms and dinoflagellates. Helgoländer Wissenschaftliche Meeresuntersuchungen 30, 192–200. https://doi.org/10.1007/BF02207835

EPA, 2011. Pesticides Industry Sales and Usage: 2006 and 2007 Market Estimates. U.S. Environ. Prot. Agency 1–41. https://doi.org/https://www.epa.gov/sites/production/files/2015-10/documents/market_estimates2007.pdf

European Council, 2000. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 Establishing a Framework for Community Action in the Field of Water Policy. Office for official publications of the European Communities, Brussels., Office for official publications of the European Communities, Brussels.

Everaert, G., Bennetsen □, E., Goethals, P.L.M., 2016. An applicability index for reliable and applicable decision trees in water quality modelling. Ecol. Inform. 32, 1–6. https://doi.org/10.1016/j.ecoinf.2015.12.004

Fabrègue, M., 2014. Extraction d'informations synthétiques à partir de données séquentielles : application à l'évaluation de la qualité des rivières. Université de Strasbourg.

Fabrègue, M., Braud, A., Bringay, S., 2013. OrderSpan: Mining Closed Partially Ordered Patterns, in: Tucker A., Höppner F., Siebes A., Swift S. (Eds) Advances in Intelligent Data Analysis XII. IDA 2013. Lecture Notes in Computer Science, Vol 8207. Springer, Berlin, Heidelberg. Springer, Berlin, Heidelberg, pp. 186–197. https://doi.org/https://doi.org/10.1007/978-3-642-41398-8_17

Fabrègue, M., Braud, A., Bringay, S., Grac, C., Le Ber, F., Levet, D., Teisseire, M., 2014. Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. Ecol. Inform. 24, 210–221. https://doi.org/10.1016/j.ecoinf.2014.09.003

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., 1996. Advances in knowledge discovery and data mining, AAAI Press. ed.

Feio, M.J., Dolédec, S., 2012. Integration of invertebrate traits into predictive models for indirect assessment of stream functional integrity: A case study in Portugal. Ecol. Indic. 15, 236–247. https://doi.org/10.1016/j.ecolind.2011.09.039

Feld, C.K., Birk, S., Eme, D., Gerisch, M., Hering, D., Kernan, M., Maileht, K., Mischke, U., Ott, I., Pletterbauer, F., Poikane, S., Salgado, J., Sayer, C.D., Van Wichelen, J., Malard, F., 2016a. Disentangling the effects of land use and geo-climatic factors on diversity in European freshwater ecosystems. Ecol. Indic. 60, 71–83. https://doi.org/10.1016/j.ecolind.2015.06.024

Feld, C.K., Segurado, P., Gutiérrez-Cánovas, C., 2016b. Analysing the impact of multiple stressors in aquatic biomonitoring data: A 'cookbook' with applications in R. Sci. Total Environ. 573, 1320–1339. https://doi.org/10.1016/j.scitotenv.2016.06.243

Forio, M.A.E., Mouton, A., Lock, K., Boets, P., Nguyen, T.H.T., Damanik Ambarita, M.N., Musonge, P.L.S., Dominguez-Granda, L., Goethals, P.L.M., 2016a. Fuzzy modelling to identify key drivers of ecological water quality to support decision and policy making. Environ. Sci. Policy 68, 58–68. https://doi.org/10.1016/j.envsci.2016.12.004

Forio, M.A.E., Van Echelpoel, W., Dominguez-Granda, L., Mereta, S.T., Ambelu, A., Hoang, T.H., Boets, P., Goethals, P.L.M., 2016b. Analysing the effects of water quality on the occurrence of freshwater macroinvertebrate taxa among tropical river basins from different continents. AI Commun. 29, 665–685. https://doi.org/10.3233/AIC-160712

Friberg, N., Angelopoulos, N.V., Buijse, A.D., Cowx, I.G., Kail, J., Moe, T.F., Moir, H., O'Hare, M.T., Verdonschot, P.F.M., Wolter, C., 2016. Chapter Eleven – Effective River Restoration in the 21st Century: From Trial and Error to Novel Evidence-Based Approaches, Advances in Ecological Research. https://doi.org/10.1016/bs.aecr.2016.08.010

Fruget, J.-F., Centofanti, M., Dessaix, J., Olivier, J.-M., Druart, J.-C., Martinez, P.-J., 2001. Temporal and spatial dynamics in large rivers : example of a long-term monitoring of the middle Rhone River. Ann. Limnol. - Int. J. Limnol. 37, 237–251. https://doi.org/10.1051/limn/2001021

Fytilis, N., Rizzo, D.M., 2013. Coupling self-organizing maps with a Naïve Bayesian classifier: Stream classification studies using multiple assessment data. Water Resour. Res. 49, 7747–7762. https://doi.org/10.1002/2012WR013422

Gandomi, A., Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. Int. J. Inf. Manage. 35, 137–144. https://doi.org/10.1016/J.IJINFOMGT.2014.10.007

Gautier, Y., 2001. Sandoz, accident écologique de l'usine de Bâle, 31 octobre 1886, in: Michel, A. (Ed.), Dictionnaire de l'écologie. pp. 1199–1200.

Geary, R.C., 1954. The Contiguity Ratio and Statistical Mapping. Inc. Stat. 5, 115. https://doi.org/10.2307/2986645

Geng, L., Hamilton, H.J., 2006. Interestingness measures for data mining: A survey. ACM Comput. Surv. 38, 3. https://doi.org/10.1145/1132960.1132963

George, A., Binu, D., 2012. DRL-Prefixspan: A novel pattern growth algorithm for discovering downturn, revision and launch (DRL) sequential patterns. Open Comput. Sci. 2, 426–439. https://doi.org/10.2478/s13537-012-0030-8

Gibert, K., Izquierdo, J., Holmes, G., Athanasiadis, I., Comas, J., Sanchez-Marre, M., 2008. On the role of pre and post-processing in environmental data mining Volume 3, 1937–1958.

Gibert, K., Izquierdo, J., Sànchez-Marrè, M., Hamilton, S.H., Rodríguez-Roda, I., Holmes, G., 2018. Which method to use? An assessment of data mining methods in Environmental Data Science. Environ. Model. Softw. 110, 3–27. https://doi.org/10.1016/j.envsoft.2018.09.021

Girard, C., Rinaudo, J.-D., Pulido-Velazquez, M., Caballero, Y., 2015. An interdisciplinary modelling framework for selecting adaptation measures at the river basin scale in a global change scenario. Environ. Model. Softw. 69, 42–54. https://doi.org/10.1016/j.envsoft.2015.02.023

Giraudel, J.L., Lek, S., 2001. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. Ecol. Modell. 146, 329–339. https://doi.org/10.1016/S0304-3800(01)00324-6

Grac, C., Herrmann, A., Le Ber, F., Trémolières, M., Braud, A., Handja, A., Lachiche, N., 2006. Mining a database on Alsatian rivers. Fr. Res. Publ. 2263–2270.

Hamylton, S., 2013. Five practical uses of spatial autocorrelation for studies of coral reef ecology. Mar. Ecol. Prog. Ser. 478, 15–25. https://doi.org/10.3354/meps10267

Haury, J., Peltre, M.-C., Trémolières, M., Barbe, J., Thiébaut, G., Bernez, I., Daniel, H., Chatenet, P., Haan-Archipof, G., Muller, S., Dutartre, A., Laplace-Treyture, C., Cazaubon, A., Lambert-Servien, E., 2006. A new method to assess water trophy and organic pollution - The Macrophyte Biological Index for Rivers (IBMR): Its application to different types of river and pollution. Hydrobiologia 570, 153–158. https://doi.org/10.1007/s10750-006-0175-3

Hering, D., Borja, A., Carstensen, J., Carvalho, L., Elliott, M., Feld, C.K., Heiskanen, A.S., Johnson, R.K., Moe, J., Pont, D., Solheim, A.L., de Bund, W. van, 2010. The European Water Framework Directive at the age of 10: A critical review of the achievements with recommendations for the future. Sci. Total Environ. 408, 4007–4019. https://doi.org/10.1016/j.scitotenv.2010.05.031

Hering, D., Moog, O., Sandin, L., Verdonschot, P.F.M., 2004. Overview and Application of the AQEM Assessment System. Hydrobiologia 516, 1–20. https://doi.org/10.1023/B:HYDR.0000025255.70009.a5

Honda, G., 2019. Caractérisation temporelle de l'état biologique des rivières en fonction des pressions physico-chimiques. Rapport de fin d'étude pour l'obtention du diplôme d'Ingénieur de l'ENGEES, Strasbourg (direction Grac et Le Ber).

Illies, J., Botosaneanu, L., 1963. Problèmes et méthodes de classification et de la zonation écologique des eaux courantes, considérées surtout du point de vue faunistique. Int. Vereinigung für Theor. und Angew. Limnol. 12, 1–57.

Imen, S., Chang, N. Bin, Yang, Y.J., 2015. Developing the remote sensing-based early warning system for monitoring TSS concentrations in Lake Mead. J. Environ. Manage. 160, 73–89. https://doi.org/10.1016/j.jenvman.2015.06.003

IPBES, 2019. Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science- Policy Platform on Biodiversity and Ecosystem Services.

Jaccard, P., 1901. Etude comparative de la distribution florale dans une portion des Alpes et du Jura. Bull. la société Vaudoise des Sci. Nat. 37, 547–579.

Johnson, L.B., Host, G.E., 2010. Recent developments in landscape approaches for the study of aquatic ecosystems. J. North Am. Benthol. Soc. 29, 41–66. https://doi.org/10.1899/09-030.1

Johnson, R.K., Hering, D., Furse, M.T., Verdonschot, P.F.M., 2006. Indicators of ecological change: comparison of the early response of four organism groups to stress gradients. Hydrobiologia 566, 139–152. https://doi.org/10.1007/s10750-006-0113-4

Kail, J., Guse, B., Radinger, J., Schröder, M., Kiesel, J., Kleinhans, M., Schuurman, F., Fohrer, N., Hering, D., Wolter, C., 2015. A Modelling Framework to Assess the Effect of Pressures on River Abiotic Habitat Conditions and Biota. PLoS One 1–21. https://doi.org/10.5061/dryad.dq87d.River

Kristensen, P., 2004. The DPSIR Framework.UNEP Headquarters, on a comprehensive/detailed assessment of the vulnerability of water ressources to environmental change in Africa using basin approach. Nairobi, Kenya.

Lafont, M., 2001. A conceptual approach to the biomonitoring of freshwater: The Ecological Ambience System. J. Limnol. 60, 17–24. https://doi.org/10.4081/jlimnol.2001.s1.17

Lafont, M., Camus, J.C., Fournier, A., Sourp, E., 2001. A practical concept for the ecological assessment of aquatic ecosystems: Application on the River Dore in France. Aquat. Ecol. 35, 195–205. https://doi.org/10.1023/A:1011413806318

Lalande, N., 2013. Impacts multi-échelles de l'occupation du sol sur l'état écologique des cours d'eau - Elaboration et test d'un cadre d'analyse et de modélisation. Thèse. AgroParisTech Montpellier.

Lalande, N., Cernesson, F., Decherf, A., Tournoud, M.-G., 2014. Implementing the DPSIR framework to link water quality of rivers to land use: methodological issues and preliminary field test. Int. J. River Basin Manag. 5124, 1–17. https://doi.org/10.1080/15715124.2014.906443

Landuyt, D., Broekx, S., Engelen, G., Uljee, I., Van der Meulen, M., Goethals, P.L.M., 2016a. The importance of uncertainties in scenario analyses - A study on future ecosystem service delivery in Flanders. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2016.02.098

Landuyt, D., Broekx, S., Goethals, P.L.M., 2016b. Bayesian belief networks to analyse trade-offs among ecosystem services at the regional scale. Ecol. Indic. 71, 327–335. https://doi.org/10.1016/j.ecolind.2016.07.015

Laney, D., 2001. 3D data management: Controlling data volume, velocity and variety.

Larras, F., Coulaud, R., Gautreau, E., Billoir, E., Rosebery, J., Usseglio-Polatera, P., 2017. Assessing anthropogenic pressures on streams: A random forest approach based on benthic diatom communities. Sci. Total Environ. xxx. https://doi.org/10.1016/j.scitotenv.2017.02.096

Larsen, S., Mancini, L., Pace, G., Scalici, M., Tancioni, L., 2012. Weak Concordance between Fish and Macroinvertebrates in Mediterranean Streams. PLoS One 7. https://doi.org/10.1371/journal.pone.0051115

Legay, J.., 1999. L'evaluation scientifique d'objets de recherche complexes releve-t-elle d'une situation epistemologique nouvelle ? Natures Sci. SocWs 7, 60–64.

Lek, S., Belaud, A., Dimopoulos, I., Lauga, J., Moreau, J., 1995. Improved estimation, using neural networks, of the food consumption of fish populations. Mar. Freshw. Res. 46, 1229–1236. https://doi.org/10.1071/MF9951229

Lek, S., Guiresse, M., Giraudel, J.-L., 1999. Predicting stream nitrogen concentration from watershed features using neural networks. Water Res. 33, 3469–3478. https://doi.org/10.1016/S0043-1354(99)00061-5

Lemm, J.U., Feld, C.K., 2017. Identification and interaction of multiple stressors in central European lowland rivers. Sci. Total Environ. 603–604, 148–154. https://doi.org/10.1016/j.scitotenv.2017.06.092

Lichstein, J.W., Simons, T.R., Shriner, S.A., Franzreb, K.E., 2002. Spatial autocorrelation and autoregression models in ecology. Ecol. Monogr. 72, 445–463. https://doi.org/10.1890/0012-9615(2002)072[0445:SAAAMI]2.0.CO;2

Liess, M., Beketov, M.A., 2011. Traits and stress: Keys to identify community effects of low levels of toxicants in test systems. Ecotoxicology 20, 1328–1340. https://doi.org/10.1007/s10646-011-0689-y

Malaj, E., Von Der Ohe, P.C., Grote, M., Kühne, R., Mondy, C.P., Usseglio-Polatera, P., Brack, W., Schäfer, R.B., 2014. Organic chemicals jeopardize the health of freshwater ecosystems on the continental scale. Proc. Natl. Acad. Sci. U. S. A. 111, 9549–9554. https://doi.org/10.1073/pnas.1321082111

Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 209–220.

Markus, M., Hejazi, M.I., Bajcsy, P., Giustolisi, O., Savic, D.A., 2010. Prediction of weekly nitrate-N fluctuations in a small agricultural watershed in Illinois. J. Hydroinformatics 12, 251. https://doi.org/10.2166/hydro.2010.064

Marzin, A., 2013. Ecological assessment of running water using bio-indicator : associated variability and uncertainty. Thesis. AgroParisTech, Paris.

Marzin, A., Archaimbault, V., Belliard, J., Chauvin, C., Delmas, F., Pont, D., 2012. Ecological assessment of running waters: Do macrophytes, macroinvertebrates, diatoms and fish show similar responses to human pressures? Ecol. Indic. 23, 56–65. https://doi.org/10.1016/j.ecolind.2012.03.010

Meador, M.R., Goldstein, R.M., 2003. Assessing water quality at large geographic scales: Relations among land use, water physicochemistry, riparian condition, and fish community structure. Environ. Manage. 31, 504–517. https://doi.org/10.1007/s00267-002-2805-5

MEDD, AE, 2003. Système d'évaluation de la qualité de l'eau des cours d'eau (SEQ-Eau), version 2 - étude Inter-Agences N°52 réalisée par le Ministère de l'Ecologie et du Développement Durable (MEDD) et les Agences de l'Eau (AE), France.

MEEDDAT, 2009. Guide Technique Evaluation de l'état des eaux douces de surface de métropole édité par le Ministère de l'Ecologie, de l'Energie, du Développement Durable et de l'Aménagement du Territoire (MEEDDAT), France. Paris.

MEEM, 2016. Guide technique Relatif à l'évaluation de l'état des eaux de surfaces continentales (cours d'eau, canaux, plans d'eau) édité par le Ministère de l'Environnement, de l'Energie, et de la Mer (MEEM), France. paris.

MEEM, 2012. Guide technique Evaluation de l'état des eaux de surfaces continentales (Cours d'eau, Canaux, Plans d'eau) édité par le Ministère de l'Environnement de l'Energie et de la Mer, France.

Meyer, A., Combroux, I., Schmitt, L., Trémolières, M., 2013. Vegetation dynamics in side-channels reconnected to the Rhine River: what are the main factors controlling communities trajectories after restoration? Hydrobiologia 714, 35–47. https://doi.org/10.1007/s10750-013-1512-y

Millenium Ecosystem Assessment Programm, 2005. Ecosystems and human well-being : synthesis, Island Pre. ed.

Mondy, C.P., Usseglio-Polatera, P., 2013. Using conditional tree forests and life history traits to assess specific risks of stream degradation under multiple pressure scenario. Sci. Total Environ. 461–462, 750–760. https://doi.org/10.1016/j.scitotenv.2013.05.072

Mondy, C.P., Villeneuve, B., Archaimbault, V., Usseglio-polatera, P., 2012. A new macroinvertebrate-based multimetric index (I 2 M 2) to evaluate ecological quality of French wadeable streams fulfilling the WFD demands : A taxonomical and trait approach. Ecol. Indic. 18, 452–467. https://doi.org/10.1016/j.ecolind.2011.12.013

Moran, P.A.P., 1950. A Test for the Serial Independence of Residuals. Biometrika 37, 178. https://doi.org/10.2307/2332162

Motelay-Massei, A., Ollivon, D., Garban, B., Teil, M.J., Blanchard, M., Chevreuil, M., 2004. Distribution and spatial trends of PAHs and PCBs in soils in the Seine River basin, France. Chemosphere 55, 555–565. https://doi.org/10.1016/J.CHEMOSPHERE.2003.11.054

Nica, C., 2017. Exploring Sequential Data with Relational Concept Analysis. Thesis. Université de Strasbourg.

Nikoo, M.R., Karimi, A., Kerachian, R., Poorsepahy-Samian, H., Daneshmand, F., 2013. Rules for Optimal Operation of Reservoir-River-Groundwater Systems Considering Water Quality Targets: Application of M5P Model. Water Resour. Manag. 27, 2771–2784. https://doi.org/10.1007/s11269-013-0314-3

Nisbet, M., Verneaux, J., 1970. Composantes chimiques des eaux courantes: discussion et proposition de classes en tant que bases d'interprétation des analyses chimiques. Ann. Limonologie 6, 161–190. https://doi.org/10.1051/limn/1970015

Oberdorff, T., Hughes, R.M., 1992. Modification of an Index of Biotic Integrity Based on Fish Assemblages to Characterize Rivers of the Seine Basin ,. Hydrobiologia 228, 117–130. https://doi.org/10.1007/BF00006200

Osborne, L.L., Kovacic, D.A., 1993. Riparian vegetated buffer strips in water□quality restoration and stream management. Freshw. Biol. 29, 243–258. https://doi.org/10.1111/j.1365-2427.1993.tb00761.x

Paasche, Ø., Österblom, H., 2019. Unsustainable Science. One Earth 1, 39–42. https://doi.org/10.1016/j.oneear.2019.08.011

Piggott, J.J., Townsend, C.R., Matthaei, C.D., 2015. Reconceptualizing synergism and antagonism among multiple stressors. Ecol. Evol. 5, 1538–1547. https://doi.org/10.1002/ece3.1465

Pohl, C., 2005. Transdisciplinary collaboration in environmental research. Futures. https://doi.org/10.1016/j.futures.2005.02.009

Poor, C.J., Ullman, J.L., 2010. Using Regression Tree Analysis to Improve Predictions of Low-Flow Nitrate and Chloride in Willamette River Basin Watersheds. Environ. Manage. 46, 771–780. https://doi.org/10.1007/s00267-010-9550-y

Pringle, C.M., 1990. Nutrient spatial heterogeneity: effects on community structure, physiognomy, and diversity of stream algae. Ecology 71, 905–920. https://doi.org/10.2307/1937362

Rasmussen, J.J., McKnight, U.S., Loinaz, M.C., Thomsen, N.I., Olsson, M.E., Bjerg, P.L., Binning, P.J., Kronvang, B., 2013. A catchment scale evaluation of multiple stressor effects in headwater streams. Sci. Total Environ. 442, 420–431. https://doi.org/10.1016/j.scitotenv.2012.10.076

Reid, W. V., Mooney, H.A., 2016. The Millennium Ecosystem Assessment: Testing the limits of interdisciplinary and multi-scale science. Curr. Opin. Environ. Sustain. https://doi.org/10.1016/j.cosust.2015.11.009

Ren, J., Wang, L., Dong, J., Hu, C., Wang, K., 2009. A Novel Sequential Pattern Mining Algorithm for the Feature Discovery of Software Fault, in: 2009 International Conference on Computational Intelligence and Software Engineering. IEEE, pp. 1–4. https://doi.org/10.1109/CISE.2009.5367106

Reyjol, Y., Argillier, C., Bonne, W., Borja, A., Buijse, A.D., Cardoso, A.C., Daufresne, M., Kernan, M., Ferreira, M.T., Poikane, S., Prat, N., Solheim, A.L., Stroffek, S., Usseglio-Polatera, P., Villeneuve, B., van de Bund, W., 2014. Assessing the ecological status in the context of the European Water Framework Directive: Where do we go now? Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2014.07.119

Reyjol, Y., Tedesco, P.A., Lim, P., 2008. Stage-dependent spatial synchrony revealed for fish populations in the Garonne River (SW France). Aquat. Sci. 70, 179–185. https://doi.org/10.1007/s00027-008-8030-4

Robach, F., Thiébaut, G., Trémolières, M., Muller, S., 1996. A reference system for continental running waters: Plant communities as bioindicators of increasing eutrophication in alkaline and acidic waters in north-east France, in: Hydrobiologia. Kluwer Academic Publishers, pp. 67–76. https://doi.org/10.1007/BF00012736

Rodela, R., Alasevic, D., 2017. Crossing disciplinary boundaries in environmental research: Interdisciplinary engagement across the Slovene research community. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2016.08.144

Sallaberry, A., Pecheur, N., Bringay, S., Roche, M., Teisseire, M., 2011. Sequential patterns mining and gene sequence visualization to discover novelty from microarray data. J. Biomed. Inform. 44, 760–774. https://doi.org/10.1016/J.JBI.2011.04.002

Schäfer, R.B., Von Der Ohe, P.C., Rasmussen, J., Kefford, B.J., Beketov, M.A., Schulz, R., Liess, M., 2012. Thresholds for the effects of pesticides on invertebrate communities and leaf breakdown in stream ecosystems. Environ. Sci. Technol. https://doi.org/10.1021/es2039882

Schulz, R., Liess, M., 1999. A field study of the effects of agriculturally derived insecticide input on stream macroinvertebrate dynamics. Aquat. Toxicol. 46, 155–176. https://doi.org/10.1016/S0166-445X(99)00002-8

Serrano Balderas, E.C., Berti-Equille, L., Hernández, M.A.A., Grac, C., 2017. Principled data preprocessing: Application to biological aquatic indicators of water pollution. Proc. - Int. Work. Database Expert Syst. Appl. DEXA 2017-Augus, 52–56. https://doi.org/10.1109/DEXA.2017.27

Sihtmäe, M., Blinova, I., Künnis-Beres, K., Kanarbik, L., Heinlaan, M., Kahru, A., 2013. Ecotoxicological effects of different glyphosate formulations. Appl. Soil Ecol. 72, 215–224. https://doi.org/10.1016/j.apsoil.2013.07.005

Staentzel, C., Arnaud, F., Combroux, I., Schmitt, L., Trémolières, M., Grac, C., Piégay, H., Barillier, A., Chardon, V., Beisel, J.-N., 2017. How do instream flow increase and gravel augmentation impact biological communities in large rivers: A case study on the Upper Rhine River. River Res. Appl. https://doi.org/10.1002/rra.3237

Statzner, B., Bêche, L.A., 2010. Can biological invertebrate traits resolve effects of multiple stressors on running water ecosystems? Freshw. Biol. https://doi.org/10.1111/j.1365-2427.2009.02369.x

Stehle, S., Schulz, R., 2015. Agricultural insecticides threaten surface waters at the global scale. Proc. Natl. Acad. Sci. U. S. A. 112, 5750–5755. https://doi.org/10.1073/pnas.1500232112

Stendera, S., Adrian, R., Bonada, N., Cañedo-Argüelles, M., Hugueny, B., Januschke, K., Pletterbauer, F., Hering, D., 2012. Drivers and stressors of freshwater biodiversity patterns across different ecosystems and scales: a review. Hydrobiologia 696, 1–28. https://doi.org/10.1007/s10750-012-1183-0

Stoddard, J.L., Jeffries, D.S., Lükewille, A., Clair, T.A., Dillon, P.J., Driscoll, C.T., Forsius, M., Johannessen, M., Kahl, J.S., Kellogg, J.H., Kemp, A., Mannlo, J., Monteith, D.T., Murdoch, P.S., Patrick, S., Rebsdorl, A., Skjelkvale, B.L., Stainton, M.P., Traaen, T., Van Dam, H., Webster, K.E., Wleting, J., Wllander, A., 1999. Regional trends in aquatic recovery from acidification in North America and Europe. Nature 401, 575–578. https://doi.org/10.1038/44114

Strahler, A.N., 1957. Quantitative analysis of watershed geomorphology. Trans. Am. Geophys. Union 38.

Tamisier, V., Bilodeau, C., Thommeret, N., Kreutzenberger, K., Melun, G., 2014. Caractérisation hydromorphologique des cours d'eau français (Carhyce). Valorisation des données Carhyce pour la construction d'un outil d'aide à la gestion des cours d'eau. Rapport scientifique CNRS (LGP-LADYSS)/Université de Paris Panthéon-Sorbonne/ESGT/AFB.

Teil, M.-J., Blanchard, M., Chevreuil, M., 2004. Atmospheric deposition of organochlorines (PCBs and pesticides) in northern France. Chemosphere 55, 501–514. https://doi.org/10.1016/J.CHEMOSPHERE.2003.11.064

Thiébaut, G., 2006. Aquatic macrophyte approach to assess the impact of disturbances on the diversity of the ecosystem and on river quality. Int. Rev. Hydrobiol. 91, 483–497. https://doi.org/10.1002/iroh.200610868

Thioulouse, J., Chessel, D., Doledec, S., Olivier, J.M., 1997. ADE-4: A multivariate analysis and graphical display software. Stat. Comput. 7, 75–83. https://doi.org/10.1023/A:1018513530268

Thorndike, R.L., 1953. Who belongs in the family? Psychometrika 18, 267–276.

Trémolières, M., 1994. Changes in aquatic vegetation in Rhine floodplain streams in Alsace in relation to disturbance. J. Veg. Sci. 5, 169–178. https://doi.org/doi: 10.2307/3236149

Tsai, W.-P., Huang, S.-P., Cheng, S.-T., Shao, K.-T., Chang, F.-J., 2016. A data-mining framework for exploring the multi-relation between fish species and water quality through self-organizing map. Sci. Total Environ. 579, 474–483. https://doi.org/10.1016/j.scitotenv.2016.11.071

Usseglio-Polatera, P., Bournaud, M., Richoux, P., Tachet, H., 2000. Biological and ecological traits of benthic freshwater macroinvertebrates: Relationships and definition of groups with similar traits. Freshw. Biol. 43, 175–205. https://doi.org/10.1046/j.1365-2427.2000.00535.x

Van Dam, H., Mertens, A., Sinkeldam, J., 1994. A coded checklist and ecological indicator values of freshwater diatoms from The Netherlands. Netherlands J. Aquat. Ecol. 28, 117–133. https://doi.org/10.1007/BF02334251

Van Looy, K., Piffady, J., Tormos, T., Villeneuve, B., Valette, L., Chandesris, A., Souchon, Y., 2015. Unravelling River System Impairments in Stream Networks with an Integrated Risk Approach. Environ. Manage. 55, 1343–1353. https://doi.org/10.1007/s00267-015-0477-1

Van Urk, G., Kerkum, F., Van Leeuwen, C.J., 1993. Insects and insecticides in the Lower Rhine. Water Res. 27, 205–213. https://doi.org/10.1016/0043-1354(93)90077-U

Verdonschot, P.F.M., Nijboer, R.C., 2004. Testing the European stream typology of the Water Framework Directive for macroinvertebrates. Hydrobiologia 516, 35–54. https://doi.org/10.1023/B:HYDR.0000025257.30311.b7

Verneaux, J., 1973. Recherches écologiques sur le réseau hydrographique du Doubs : essai de biotypologie. Thèse. Université de Franche-Comté, Besançon.

Villeneuve, B., 2016. Modèles multi-stress et multi-échelles de l'état écologique : vers une analyse du risque d'altération des cours d'eau et des bassins versants. Thèse IRSTEA. Université de Lorraine, Metz.

Villeneuve, B., Piffady, J., Valette, L., Souchon, Y., Usseglio-Polatera, P., 2018. Direct and indirect effects of multiple stressors on stream invertebrates across watershed, reach and site scales: A structural equation modelling better informing on hydromorphological impacts. Sci. Total Environ. 612, 660–671. https://doi.org/10.1016/J.SCITOTENV.2017.08.197

Villeneuve, B., Souchon, Y., Usseglio-Polatera, P., Ferréol, M., Valette, L., 2015. Can we predict biological condition of stream ecosystems? A multi-stressors approach linking three biological indices to physico-chemistry, hydromorphology and land use. Ecol. Indic. https://doi.org/10.1016/j.ecolind.2014.07.016

Waite, I.R., Kennen, J.G., May, J.T., Brown, L.R., Cuffney, T.F., Jones, K.A., Orlando, J.L., 2014. Stream macroinvertebrate response models for bioassessment metrics: Addressing the issue of spatial scale. PLoS One 9. https://doi.org/10.1371/journal.pone.0090944

Walley, W., Džeroski, S., 1996. Biological monitoring: A comparison between Bayesian, neural and machine learning methods of water quality classification, in: Environmental Software Systems: Proceedings of the International Symposium on Environmental Software Systems. pp. 229–240. https://doi.org/10.1007/978-0-387-34951-0_20

Wasson, J.-G., 2001. Les questions de recherche posées par la Directive Cadre Européenne sur l ' Eau : problématique pour les eaux de surface continentales Research questions arising from the European Water Frame work Directive : topics related to inland surface waters 1, 1–19.

Wasson, J.-G., Chandesris, A., Pella, H., Blanc, L., 2004. Les hydro-écorégions: une approche fonctionnelle de la typologie des rivières pour le Directive cadre européenne sur l'eau. Ingénieries - E A T 40, 3–10.

Wasson, J., Chandesris, A., Pella, H., Blanc, L., 2004. Les hydro-écorégions: une approche fonctionnelle de la typologie des rivières pour la Directive cadre européenne sur leau. Ingénieries.

Webb, B.W., 1996. Trends in stream and river temperature. Hydrol. Process. 10, 205–226. https://doi.org/10.1002/(SICI)1099-1085(199602)10:2<205::AID-HYP358>3.0.CO;2-1

Webb, B.W., Hannah, D.M., Moore, R.D., Brown, L.E., Nobilis, F., 2008. Recent advances in stream and river temperature research. Hydrol. Process. https://doi.org/10.1002/hyp.6994

Willby, N.J., Abernethy, V.J., Demars, B.O.L., 2000. Attribute-based classification of European hydrophytes and its relationship to habitat utilization. Freshw. Biol. 43, 43–74. https://doi.org/10.1046/j.1365-2427.2000.00523.x

Winemiller, K.O., Rose, K.A., 1992. Patterns of life-history diversification in North American fishes: implications for population regulation. Can. J. Fish. Aquat. Sci. 49, 2196–2218. https://doi.org/10.1139/f92-242

Wogram, J., Liess, M., 2001. Rank ordering of macroinvertebrate species sensitivity to toxic compounds by comparison with that of Daphnia magna. Bull. Environ. Contam. Toxicol. 67, 360–367. https://doi.org/10.1007/s00128-001-0133-8

Xue, D., De Baets, B., Van Cleemput, O., Hennessy, C., Berglund, M., Boeckx, P., 2013. Classification of Nitrate Polluting Activities through Clustering of Isotope Mixing Model Outputs. J. Environ. Qual. 42, 1486. https://doi.org/10.2134/jeq2012.0456

PARCOURS PROFESSIONNEL

Situation

50 ans, vie maritale, 2 enfants

✉ ADRESSE PERSONNELLE
30, route des Basses-Huttes
68370 ORBEY

✉ ADRESSE PROFESSIONNELLE
ENGEES, 1, quai Koch
67070 Strasbourg

Formation académique & concours

1999 : **Ingénieur des Travaux Ruraux** par concours direct au Ministère de l'Agriculture et de la Pêche.

1992 : **DESS d'HYDROBIOLOGIE** " Eaux continentales, pollutions et aménagements ". Université de Franche-Comté – (Resp. J. VERNEAUX), Besançon (25) ;

Stage de fin détude (6 mois) : « Caractérisation physico-chimique de quatre lacs du Jura » sous la direction de J. Verneaux, Université de Franche-Comté.

1991: Maitrise « Sciences de l'Environnement et Chimie analytique », Université d'Aix-Marseille, Marseille (13) ;

Expérience professionnelle : Hydroécologue, Ingénieur Environnement et Agriculture (IAE – Ministère de l'Agriculture)

8 ans en service opérationnel en Outre-Mer : création des 1ers réseaux de surveillance des rivières réunionnaises puis guadeloupéennes

1994 à 1999: **Ingénieur Hydrobiologiste** à l'Observatoire Réunionnais de l'Eau, à La Réunion :

- ⇨ création et gestion du réseau de surveillance de la qualité des eaux,
- ⇨ inventaire floristiques et faunistiques (premier atlas des invertébrés réunionnais),
- ⇨ surveillances physico-chimiques automatisées, études spécifiques en qualité,
- ⇨ interlocuteur "qualité des eaux" des différents partenaires (état, collectivités, privés, ...),
- ⇨ formatrice ponctuelle (responsable du module hydrobiologie en maîtrise d'écologie et environnement de l'Université de la Réunion).

1999 à 2003: **Ingénieur Travaux Ruraux « Chargée de mission en hydrobiologie et qualité des milieux aquatiques »** au Service Eau, Milieux Aquatiques et Déchets (SEMAD) de la Direction Régionale de l'Environnement (DIREN) de Guadeloupe :

- ⇨ assistance au chef de service dont intérim du service, encadrement,
- ⇨ création et gestion du réseau de surveillance hydrobiologique des eaux de l'archipel,
- ⇨ conduite et accompagnement d'étude et de projets, développement de la base de données qualité,
- ⇨ définition d'indices biologiques (invertébrés et diatomées) en partenariats universitaires,
- ⇨ communications orales (tous public) et écrites (plaquette, cartes de qualité, sites web …),

16 ans en poste d'enseignant-chercheur à l'ENGEES (Ecole Nationale du Génie de l'Eau et de l'Environnement de Strasbourg)

2003 -2019 : **Ingénieur Agriculture Environnement (dont 5 ans à 60% et 8 ans à 80%) rattachée en 2002-08 au CEVH** (Centre d'Ecologie Végétale et d'Hydrologie, UMR Université de Strasbourg –UDS-/ENGEES), **puis de 2009-13, au LHYGES** (Laboratoire d'Hydrologie et de Géochimie de Strasbourg UMR UDS-CNRS-ENGEES), et **depuis 2013** au **LIVE** (Laboratoire Image, Ville et Environnement, UMR UDS-CNRS-ENGEES).

- ⇨ élaborations et conduites, ou collaborations à des projets de recherche sur les invertébrés aquatiques, l'évaluation de l'état des milieux aquatiques, les liens pressions-impacts, les suivis pré-&post-restaurations, (financements : ANR, AFB, Agence de l'Eau, DREAL, ONF, Collectivités), - voir détail ci-après ;
- ⇨ enseignements :
  - o volume : ~120 heures/ an
  - o niveau : L3, M1 &2 ;
  - o formations : ingénieurs, masters, mastères & licence professionnelle, formation continue
  - o matières : écologie, hydroécologie, bio-indication, chimie, surveillance, gestion, aménagement et restauration des milieux aquatiques,
  - o responsabilité de 3 Unités d'Enseignements,
- ⇨ ingénierie pédagogique et vulgarisation scientifique (organisation journées Scientifiques et Techniques de rencontre entre professionnels de l'eau et chercheurs, 2010, 2014),
- ⇨ membre de la Commission de Normalisation AFNOR T95-F : indicateurs biologiques aquatiques,
- ⇨ encadrements de stages de niveau master (12 depuis 2003) et licence (18 depuis 2003),
- ⇨ encadrements d'apprentis (2 depuis 2003) de niveau master 2 (niveau master)
- ⇨ participation à l'encadrement de thèses : - voir détail ci-après.

Principaux projets

En évaluation

Projet motifs temporels de changement d'état (2019-2020):

Porteuse du projet : **Corinne Grac**
Sujet : Elaboration d'un dispositif d'extraction de motifs temporels d'altérations lors d'un changement d'état biologique, à l'échelle des hydro-éco-régions
Personnes associées: Xavier Dolques (ENGEES, ICube), Florence Le Ber (ENGEES, ICube) & Agnès Braud (Unistra, Icube)
Financement: AFB

Projet extension nationale de FRESQUEAU (2015-2018):

http://dataqual.engees.unistra.fr/fresqueau_nat_presentation
Porteuse du projet : **Corinne Grac**
Sujet : Application de la méthode de fouilles des motifs temporels partiellement ordonnés aux données nationales de suivis des rivières en France métropolitaine
Personnes associées: Xavier Dolques (ENGEES, ICube), Florence Le Ber (ENGEES, ICube) & Agnès Braud (Unistra, ICube)
Financement: ONEMA

Projet ANR FRESQUEAU (2011-2015): Fouille de données de suivis de cours d'eau

http://dataqual.engees.unistra.fr/fresqueau_presentation_f

Porteuse de projet : Florence Le Ber (ENGEES, ICube)
Sujet : Application des méthodes de fouille aux données eau
Co-porteuse de tâche : **Corinne Grac** & Flavie Cernesson (AgroParisTech, TETIS)

- ⇨ Collecte des données et intégration à la base de données
- ⇨ la définition des questions des thématiciens

Collaboration aux autres tâches :

- ⇨ l'élaboration du modèle conceptuel de la base de données,
- ⇨ l'élaboration de l'entrepôt de données associé et des cubes d'explorations des données
- ⇨ la préparation des données
- ⇨ l'interprétation des résultats produits à l'aide de deux méthodes de fouilles des motifs partiellement ordonnés et des treillis relationnels de concepts

Projet INDICES (2005-2010)

http://dataqual.engees.unistra.fr/projet_indices_presentation
Porteuse du projet : **Corinne Grac**
Sujet : Evaluer l'état des rivières en Plaine d'Alsace en combinant les réponses des indices biologiques (IBGN, IOBS, IPR, IBD & IBMR)
Personnes associées: Michèle Trémolières (Unistra, LIVE), Florence Le Ber (ENGEES, ICube) & Agnès Braud (Unistra, ICube), Sébastien Manné (ONEMA), Michel Lafont (Irstea), Luc Ector (CRP Lipmann)
Financement: Agence de l'Eau Rhin Meuse

En restauration

Projet WOERR annuel reconduit depuis 2012 à ce jour: Suivis des restaurations de milieux stagnants en vue de la ré-introduction de la cistude d'Europe (*Emys orbicularis*), à Lauterbourg (Alsace, France)

Co-porteuses de projet : Isabelle Combroux (Unistra, LIVE) & **Corinne Grac** (ENGEES, LIVE)
Sujet Suivis de la dynamique des peuplements invertébrés dans les milieux restaurés et interactions avec les peuplements végétaux aquatiques, en collaboration avec Isabelle Combroux (Unistra, LIVE), Florence Le Ber (ENGEES, ICube), Albin Meyer (Université de Metz, LIEC) et Frédéric Labat (Bureau d'étude Aquabio)
Etude des relations trophiques entre cistude et invertébrés en collaboration avec Jean-Yves Georges (CNRS, IPHC)
Financement: Conseil Départemental Haut Rhin, Conseil Scientifique ENGEES, CNRS

Projet RESTAURATION DU RHIN (2012-2018)

Porteurs du projet : Laurent Schmitt (Unistra, LIVE) & Jean-Nicolas Beisel (ENGEES, LIVE)
Sujet : Impacts de recharges sédimentaires dans le Rhin supérieur à l'aval du barrage de Kembs, France : suivi écologique des plantes, macro-invertébrés et poissons
Personnes associées: Cybill Staentzel, Isabelle Combroux & Michèle Trémolières (Unistra, LIVE), Laurent Schmitt & Valentin Chardon (Unistra, LIVE), Olivier Schlumberger & **Corinne Grac** (ENGEES, LIVE), Fanny Arnaud & Hervé Piégay (CNRS, EVS), Agnès Barillier & Alain Garnier (EDF)
Financement : EDF

Projet ROHRSCHOLLEN (2010-2015)

Porteur du projet : Laurent Schmitt (Unistra, LIVE)
Sujet : Restauration de la dynamique des habitats alluviaux rhénans sur l'île du Rohrschollen.
Personnes associées: Michèle Trémolières (Unistra, LIVE), David Eschbach (Unistra, LIVE), **Corinne Grac** puis Jean-Nicolas Beisel (ENGEES, LIVE),

Pascal Finaud-Guyot (ENGEES, ICube), Sylvain Weill & Sylvain Payraudeau (ENGEES, LHYGES), Gwénaël Imfeld (CNRS, LHYGES)
Financement : LIFE+, Euro-métropole de Strasbourg, IDEX, Conseil Scientifique de l'ENGEES

Participations à l'encadrement de 6 thèses

Nadia Fernandez, abandonnée en 2017 (évaluations restaurations rivières)
"Indicateurs de restauration d'écosystèmes aquatiques basés sur les invertébrés"
Université de Strasbourg, spécialité Sciences de la Terre et de l'Environement.
Financement: thèse CIFRE avec le Bureau d'étude Aquabio.
Directeur: Jean-Nicolas Beisel; co-encadrement: Corinne Grac (ENGEES, LIVE) & Frédéric Labat (Aquabio).

Cristina Nica, **2017** (Fouille de données)

« Exploring sequential data with relational concept analysis »
Université de Strasbourg, spécialité informatique.
Directrice : Florence Le Ber (ENGEES, ICube); co-encadrement Agnès Braud (Unistra, ICube) & Corinne Grac (ENGEES, LIVE)

Eva Carmina Serrano Balderas, 2017 (évaluations rivières mexicaines et pré-traitements des données)

" Preprocessing and analysis of environmental data: application to the water quality assessment of Mexican rivers"
Université de Montpellier, spécialité informatique.
Financement mexicain Conacyt.
Directrice: Laure Berti-Equille (IRD,Espace DEV) , Co-Directrice: Maria Aurora Armienta Hernandez (Univ. Mexico), co-encadrants: Corinne Grac (ENGEES, LIVE) et Jean-Christophe Desconnets (IRD, Espace DEV)

Juliane Wiederkehr, 2015, (incertitudes et bioindications)

"Estimation des incertitudes associées aux indices macroinvertébrés et macrophytes pour l'évaluation de l'état écologique des cours d'eau"
Université de Strasbourg, spécialité Sciences de la Terre et de l'Environement.
Co-financement: ENGEES & Bureau d'étude Aquabio.
Directrice: Michèle Trémolières (Unistra, LIVE), co-directrice: Florence Le Ber (ENGEES, ICube), Co-encadrants: Corinne Grac (ENGEES, LIVE) & Frédéric Labat (Aquabio).

Albin Meyer, 2012 (dynamique de restauration, reconnexion au Rhin)

"Processus et dynamique de la recolonisation et de la biodiversité dans les bras du Rhin et autres cours d'eau restaurés de la Plaine d'Alsace après reconnexion"
Université de Strasbourg, spécialité Sciences de la Terre et de l'Environement.
Financement: Unistra

Directrice: Michèle Trémolières (Unistra, LIVE), Co-encadrants: Isabelle Combroux (Unistra, LIVE) & Corinne Grac (ENGEES, LIVE)

Aurélie Bertaux, 2010 (Informatique appliquée à l'hydroécologie)

"Treillis de Galois pour les contextes multi-valués flous : Application à l'étude des traits de vie en hydrobiologie"
Université de Strasbourg, spécialité informatique.
Co-financement: Agence de l'Eau Rhin-Meuse & Région Alsace.
Directrice: Florence Le Ber (ENGEES, ICube), co-directrice: Michèle Trémolières (Unistra, LIVE), co-encadrantes: Corinne Grac (ENGEES, LIVE) & Agnès Braud (ENGEES, ICube)

Valorisations scientifiques

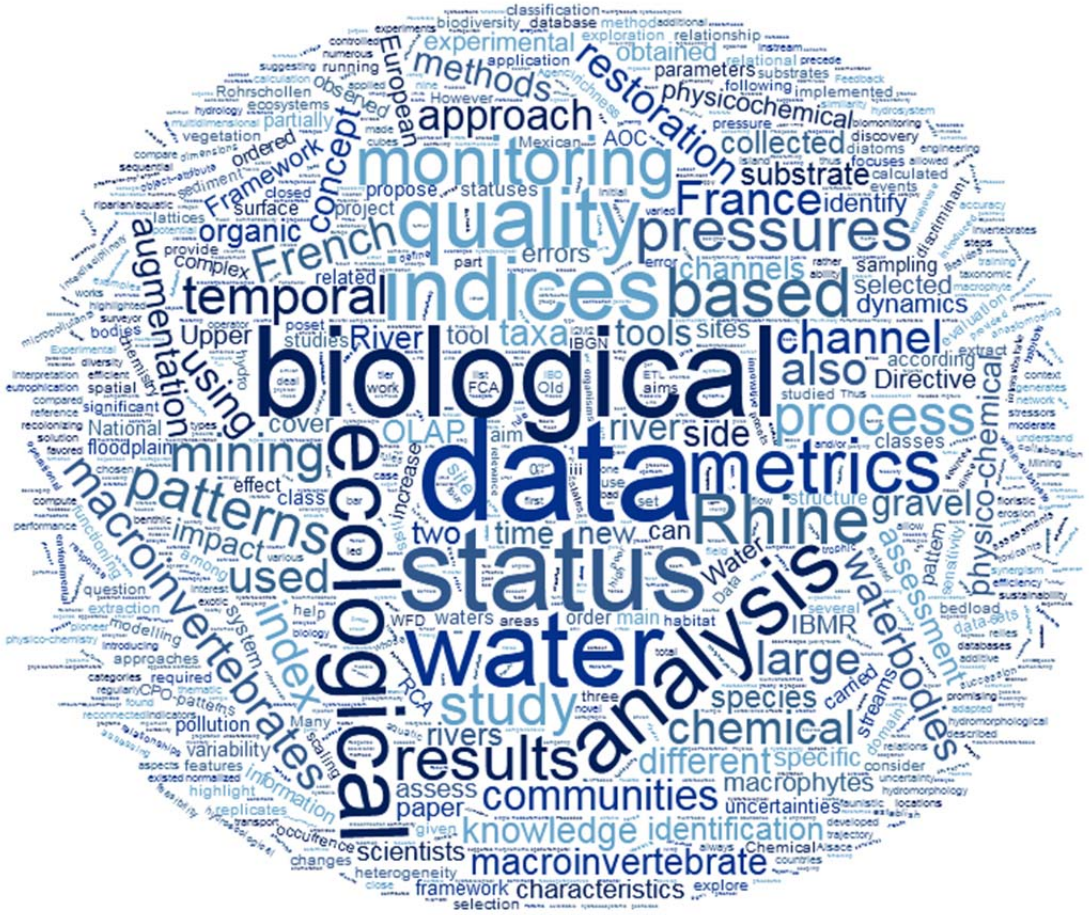


Figure 62 : nuages de mots réalisé à partir des résumés des articles scientifiques où je figure parmi les auteurs (fait par C. Staentzel)

20 Articles, dont 3 en préparation ou soumission

Braud, A., Gançarski, P., **Grac, C**., Le Ber, F., **2019**. Classification de séries temporelles hétérogènes pour le suivi de l'état des cours d'eau. Conférence Francoph. EGC - Extr. Gest. des Connaissances-, 27-31/01/2020, Bruxelles, B, en soumission.

Grac, C., Cernesson, F., Dolques, X., Braud, A., Herrmann, A., Labat, F., Teisseire, M., Trémolières, M., Le Ber, F., **2019**. Which data mining method to use for the evaluation of river ecological status – Feedback on a close collaboration between data scientists and hydro-scientists. *Enviromental Modelling and Software. En soumission.*

Grac, C., Dolques, X., Braud, A., Trémolières M., Beisel, J-N., Le Ber, F., **2019**. Mining the sequential patterns of water quality preceding biological status of waterbodies. En preparation.

Staentzel, C., Combroux, I., Barillier, A., **Grac, C**., Chanez, E., Beisel, J.-N., **2019**. Effects of a river restoration project along the Old Rhine River (France-Germany): Response of macroinvertebrate communities. Ecol. Eng. 127, 114–124. https://doi.org/10.1016/J.ECOLENG.2018.10.024

Staentzel, C., Arnaud, F., Combroux, I., Schmitt, L., Trémolières, M., **Grac, C**., Piégay, H., Barillier, A., Chardon, V., Beisel, J.-N., **2017.** How do instream flow increase and gravel augmentation impact biological communities in large rivers: A case study on the Upper Rhine River. River Res. Appl. https://doi.org/10.1002/rra.3237

Dolques, X., Le Ber, F., Huchard, M., **Grac, C.**, **2016**. Performance-friendly rule extraction in large water data-sets with AOC posets and relational concept analysis. Int. J. Gen. Syst. 45, 187–210. https://doi.org/10.1080/03081079.2015.1072927

Serrano Balderas, E.C., **Grac, C.**, Berti-Equille, L., Armienta Hernandez, M.A., **2016**. Potential application of macroinvertebrates indices in bioassessment of Mexican streams. Ecol. Indic. 61, 558–567. https://doi.org/10.1016/j.ecolind.2015.10.007

Berrahou, L., Lalande, N., Serrano Balderas, E.C., Molla, G., Berti-Équille, L., Bimonte, S., Bringay, S., Cernesson, F., **Grac, C.**, Ienco, D., Le Ber, F., Teisseire, M., **2015**. A quality-aware spatial data warehouse for querying hydroecological data. Comput. Geosci. 85, 126–135. https://doi.org/10.1016/j.cageo.2015.09.012

Bimonte, S., Boulil, K., Braud, A., Bringay, S., Cernesson, F., Dolques, X., Fabrègue, M., **Grac, C.**, Lalande, N., Le Ber, F., Teisseire, M., **2015a**. Un système décisionnel pour l'analyse de la qualité des eaux de rivières. Ingénierie des systèmes d'information 20, 143–167. https://doi.org/10.3166/isi.20.3.143-167

Bimonte, S., Boulil, K., Braud, A., Bringay, S., Cernesson, F., Dolques, X., Fabrègue, M., **Grac, C.,** Lalande, N., Le Ber, F., Teisseire, M., **2015b**. A decisional system for analysing water quality of watercourses. Ing. des Syst. d'Information 20. https://doi.org/10.3166/ISI.20.3.143-167

Wiederkehr, J., **Grac, C**., Fabrègue, M., Fontan, B., Labat, F., Le Ber, F., Trémolières, M., **2015.** Experimental study of uncertainties on the macrophyte index (IBMR) based on species identification and cover. Ecol. Indic. 50, 242–250. doi:10.1016/j.ecolind.2014.10.021

Boulil, K., Le Ber, F., Bimonte, S., **Grac, C.**, Cernesson, F., **2014**. Multidimensional modeling and analysis of large and complex watercourse data: An OLAP-based solution. Ecol. Inform. 24, 90–106. https://doi.org/10.1016/j.ecoinf.2014.07.001

Fabrègue, M., Braud, A., Bringay, S., **Grac, C.**, Le Ber, F., Levet, D., Teisseire, M., **2014.** Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. Ecol. Inform. 24, 210–221. https://doi.org/10.1016/j.ecoinf.2014.09.003

Le Ber, F., Teisseire, M., Braud, A., Cernesson, F., **Grac, C.**, Poncelet, P., **2014**. The Fresqueau project: Exploit Big Data on waterways. Ing. des Syst. d'Information 19.

Le Ber, F., Teisseire, M., Braud, A., Cernesson, F., **Grac, C.**, Poncelet, P., **2012b.** Le projet Fresqueau : exploiter les données massives concernant les cours d'eau. Ingénierie des systèmes d'information 1, 9–12. https://doi.org/10.3166/ISI.22

Wiederkehr, J., **Grac, C.**, Fabrègue, M., Fontan, B., Labat, F., Le Ber, F., Trémolières, M., **2015**. Experimental study of uncertainties on the macrophyte index (IBMR) based on species identification and cover. Ecol. Indic. 50, 242–250. doi:10.1016/j.ecolind.2014.10.021

Braud, A., Nica, C., **Grac, C.**, Le Ber, F., **2011**. A lattice-based query system for assessing the quality of hydro-ecosystems, in: CEUR Workshop Proceedings.

Grac C., Braud A., Le Ber F., Tremolieres M., **2011**. Un système d'information pour le suivi et l'évaluation de la qualité des cours d'eau – Application à l'hydro-région de la plaine d'Alsace. *Ingénierie des Systèmes d'Information*, 16, x, 9-30.

Grac, C., Herrmann, A., Le Ber, F., Trémolières, M., Braud, A., Handja, A., Lachiche, N., **2006**. Mining a database on Alsatian rivers. Fr. Res. Publ. 2263–2270.

Grac C., **Herrmann** A., **Le Ber** F., **Trémolières** M., Braud A., ***Handja*** A., Lachiche N.*, 2006* [2] : Mining a database on Alsatian rivers. In: Proccedings of the *7th International Conference on Hydroinformatics, HIC 2006, Nice, France,* volume III, pp. 2263-2270.

23 Communications scientifiques, dont 11 en congrès international

Staentzel, C., Chanez, E., Dumont, S., **Grac, C**., Hardion, L., Beisel J-N., **2019**. Improving habitat quality through restoration: a key for the protection of aquatic biodiversity and its related ecosystem services. Oral presentation, ISRS 8-13/08/2019, Vienne, A.

Grac, C., Braud, A., Le Ber, F., **2018** – Extraction de motifs temporels caractéristiques des hydro-écorégions à partir de données de surveillance des rivières françaises. Communication orale, congrés de l'Association Française de Limnologie, 22-23/11/2018, Strasbourg

Grac, C., Dolques, X., Le Ber, F., Braud, A., Cernesson, F., Trémolières, M., Beisel, J.-N., **2017a**. A new data mining approach to understand the river ecological status: first large application of closed partially ordered patterns on French aquatic data. oral presentation, 10°Symposium for European Freshwater Sciences 2-7/07/2017, Olomouc, CZ.

Grac, C., Combroux, I., Rozan, A., Meyer, A., Fernadez, N., Staentzel, C., Labat, F., Le Ber, F., Levresse, F., Kern, S., Schneider, P., Georges, J-Y., **2017b** – Retours sur la restauration écologique d'une ancienne gravière et de zones humides en plaine alluviale rhénane (site du Woeer) et leurs suivis. Communication orale, Colloque franco-allemand Retour d'expérience en retauration 11-12/05/2017, Strasbourg

Meyer, A., **Grac, C**., Schmitt, L., Combroux, I., Trémolières, M., **2017b** – Impact de la redynamisation d'anciennes annexes hydrauliques sur le dynamique des macro-invertébrés benthiques et sur la végétation aquatique rivulaire. Communication orale, Colloque franco-allemand Retour d'expérience en retauration 11-12/05/2017, Strasbourg

Meyer, A., Fernadez, N., Beisel, J-N., Georges, J-Y ., Combroux, I., Labat, F., **Grac, C**., **2017b**. Macrophytes introduced in newly-dug ponds benefifts to benthic macro-invertebrates communities. oral presentation, 10°Symposium for European Freshwater Sciences 2-7/07/2017, Olomouc, CZ.

Serrano Balderas, E.C., **Grac, C**., Berti-Equille, L., Hernández, M.A.A., Desconnets, J-C.. **2017**. Evaluation of heavy metals, pesticides and emergent content in Tula river, Mexico. oral presentation, 10°Symposium for European Freshwater Sciences 2-7/07/2017, Olomouc, CZ.

Serrano Balderas, E.C., Berti-Equille, L., Hernández, M.A.A., **Grac, C., 2017.** Principled data preprocessing: Application to biological aquatic indicators of water pollution. Proc. - Int. Work. Database Expert Syst. Appl. DEXA 2017-Augus, 52–56. doi:10.1109/DEXA.2017.27

Grac, C., Dolques, X., Le Ber, F., Braud, A ., Trémolières, M., Beisel., J-N., **2016**. First large application of closed partially ordered patterns to data collected for the French national ecological assessment of rivers: toward a new approach to understand the state of aquatic ecosystems. Congrès des doctorants ED-413, 30/11/2016, Strasbourg.

Staentzel, C., Beisel, J-N., Arnaud, F., Combroux, I., **Grac, C.**, Trémolières, M., Schmitt, L.,Piégay, H.,Barillier, A.,**2016**. Ecological impacts of sediment recharges into the upper Rhine river downstream of the Kembs diversion dam: ecological monitoring of plants, macroinvertebrates and fish. Congrès SHF : «**HydroES 2016**», 16-17/03/2016, Grenoble.

Arnaud, F., Staentzel, C., Beisel, J-N., Piégay, H., **Grac C.,** Trémolières, M., Combroux, I., Schmitt, L., Barillier, A., Garnier, A., **2015**. Geomorphic and ecological monitoring of an experimental sediment reintroduction into the Rhine River downstream of the Kembs dam. Colloque ISRIVERS, 22-26/06/2015, Lyon

Eschbach D., Schmitt L., Trémolières M., Beisel, J-N., Grac C., Finaud-Guyot P., Weill, S., Payraudeau, S., Imfeld, G., **2015**. Functional restoration of a Rhine anastomosing channel: temporal trajectory, initial state, post-restoration monitoring, modelling (Upper Rhine, France, Rohrschollen island). Colloque ISRIVERS, 22-26/06/2015, Lyon

Eschbach D., Schmitt L., Trémolières M., **Grac C**., Finaud-Guyot P., Zimmermann A., Lejeune Q., **2013**. Suivi interdisciplinaire de la restauration hydro-morphologique d'une anastomose rhénane (le Bauerngrundwasser dans l'île du Rohrschollen, France): premiers résultats. Colloque Life Walphy, 15-17/10/13, Namur, B.

Wiederkehr J., Fabregue M., Fontan B., **Grac C**., Labat F., Le Ber F., Trémolières M., **2013**. Multi index assessment of streams and associated uncertainties: application to macrophytes. 10th Symposium for European Freshwater Sciences SEFR, Munster, D, 2-4/07/13.

Combroux I., *Meyer A*., **Grac C**., Trémolières M., **2010**. Vegetation dynamics after restoration of connectivity in Rhine side channels. 7th Europeen conference on Ecological Restauration, SER 2010, Avignon France, 23-27/08/2010

Grac C., Braud A., Le Ber F., Trémolières M. **2010**. Un système d'information pour le suivi et l'évaluation de la qualité des cours d'eau. Actes du 3e Atelier "Systèmes d'Information et de Décision pour l'Environnement" SIDE 2010, Congrès Inforsid, Marseille, 25-28/5/10.

Braud A., **Grac C.,** Pristavu S., Dor E., Le Ber F. **2009**. Une démarche fondée sur les treillis de Galois pour l'aide à la qualification de l'état des milieux aquatiques. SIDE 2009, Toulouse, 26/5/09. Actes du 2e Atelier "Systèmes d'Information et de Décision pour l'Environnement" - SIDE 2009, p. 95-104.

Grac C., Le Ber F., Nobelis P., Lafont M., Trémolières M. **2009**. Response of biological indices to pressures in the running waters of the Alsace floodplain (Eastern France). Towards a proposal of a new tool for assessing ecological status of waterbodies. 6th Symposium for European Freshwater Sciences SEFS, Sinaia, RO, 17-21/8/09.

Grac C., Lafont M., Le Ber F., Nobelis P., Trémolières M. **2007**. Response of biological indices to pressures in the running waters of the Alsace floodplain (Eastern France). Towards a proposal of a new tool to assess ecological status of waterbodies.5th Symposium for European Freshwater Science, SEFS, Palermo, I, 8-13/7/07

Grac C., Ehrhard J.-L., Le Ber F., Trémolières M., **2005**a: Une base de données pour l'étude de la qualité des cours d'eau alsaciens*. Journée du groupe de travail « Fouille de données complexes », 19 mai 2005, Paris.*

Grac C., Ehrhard J.-L., Trémolières M., Le Ber F., Herrmann A., **2005**b. Une base de données pour l'étude de l'état écologique des cours d'eau alsaciens dans la cadre de la DCE. *Colloque CILO, 5-8 Juillet 2005, Vaulx en Velin.*

Barthe E., **Grac** C., Brosse S., Thomas A., **2003**. Definition d'un indice invertébrés guadeloupéen. 46ème congrès annuel de l'Association Française de Limnologie, Metz.

Leitao M., **Grac** C., Unrien F., **2003**. Premier inventaire du phytoplancton de la Guadeloupe. 46ème congrès annuel de l'Association Française de Limnologie, Metz.

4 Posters en congrès international

François, M., **Grac, C**., Combroux, I., **2019** – Calico crayfish (*Faxonius immunis*) a new invasive species in France : from biological traits to preventive measures. Poster, Aquatic Biodiv. Internat. Conf.25-28/09/2019, Sibiu, RO

Georges, J-Y., Grac, C., Quintard, B., **2015**. Predatory of the European pond turtle *Emys orbicularis* on the invasive zebra mussel *Dreissena polymorpha*. Research and conservation of European herpetofauna and its enviroment. 24-25/09/2015, Latvia, LE.

Grac, C., Berrahou, L., Bimonte, S., Boulil, K., Braud, K., Bringay, S., Cernesson, F., Dolques, X., Fabrègue, M., Fontan, B., Lalande, N., Le Ber, F., Levet, D., Molla, G., Niel, J., Teisseire, M., **2015**. Une base de données pour l'évaluation et le suivi de la qualité hydrobiologique des cours d'eau. Congrès des doctorants ED-413, 30/11/2015, Strasbourg.

Meyer A., Combroux I., **Grac** C., Schmitt L., Trémolères M., **2011**. Plant and macro-invertebrate dynamics in Rhine side channels after restoration. Conference World Large Rivers, Vienna, A, 11-15/04/11

Autres productions

Transfert de connaissances

Invitation à la journée de dialogue Recherche-Gestion de l'Etablissement Public de Bassin de Loire, juin 2014, Orléans : présentation orale

Berrahou, L., Bimonte, S., Braud, A., Bringay, S.,, Boulil, K., Cernesson, F., Dolques, Fabrègue, M., Fontan, B., Grac, C.,, Hermann, A., Labat, F., Lalande, N., Levet, D., Molla, G., Niel, J., Teisseire, M., Wiederkehr, J., Le Ber, F., **2016**. **F**resQueau: un projet de recherche sur des données massives pour évaluer la qualité des cours d'eau. Le 24 juin 2016, Orléans.

Invitation à la journée technique Cistudes, avril 2014, Strasbourg

Georges, J-Y., Grac, C., Quintard, B., 2014. Tests de prédation de la moule ébrées *Dreissena polymorpha* par la cistude d'Europe *Emys orbicularis*. Journées Techniques Cistude, 4-5/02/2014, Strasbourg

Georges, J-Y., Beisel, J-N, Combroux, I., Grac, C., Hoch, D., Kern, S., Knibiely, P., Labat, F., Levresse, F., Quintard, B., Rozan, A., Schneider, P., 2014. Site d'Etude en Ecologie Globale du Woerr : approche multicritère d'un programme de relâcher de cistudes d'Europe en Alsace. Communication orale. Journées Techniques Cistude, 4-5/02/2014, Strasbourg

Co-organisation de la Journée Scientifique et Technique de l'ENGEES et de l'ASTEE (Association Scientifique et Technique pour l'Eau et l'Environnement), **2014, Strasbourg**

« Gestion des grands volumes de données rivières : structuration, explorations innovantes, et utilisations pratiques», le 20 mars 2014. 10 orateurs, une trentaine de participants. Portée régionale.

Organisation de la Journée Scientifique et Technique de l'ENGEES et de l'ASTEE (Association Scientifique et Technique pour l'Eau et l'Environnement), **2010, Strasbourg**

« Les outils de surveillance de l'état écologique des cours d'eau. Une dynamique nouvelle impulsée par la DCE », le 9 mars 2010. 10 orateurs, une centaine de participants. Portée régionale.

20 Rapports de contrats de recherche ou équivalent

Dont

Grac C., Hoareau G., Hoarau C., 1999. L'atlas des macroinvertébrés aquatiques de La Réunion – projet de l'Observatoire Réunionnais de l'Eau, de la Région Réunion et de la Direction Régionale de l'Environnement de La Réunion, *219 p. dont 88 planches illustrées.*

Université de Strasbourg | ENGEES | Image Ville Environnement | ED 413

Corinne GRAC

Fouille temporelle des indicateurs physico-chimiques et biologiques pour l'évaluation de l'état, des pressions et de la capacité de résilience des rivières

RESUME

Les données issues de la surveillance des rivières sont volumineuses, avec des relations complexes. Des méthodes de fouille de données non supervisées peuvent s'y appliquer et donner des résultats pertinents pour leur gestion, sous réserve d'une collaboration étroite entre hydroécologues et informaticiens. L'extraction de motifs partiellement ordonnés à partir de séquences temporelles de pressions physico-chimiques précédant un état biologique a été réalisée. Ces motifs temporels permettent d'identifier une partie des pressions en cause dans un état écologique dégradé ou non, de préciser l'importance de la durée des séquences avant l'évaluation de l'état biologique, d'identifier les altérations caractéristiques à l'échelle d'une hydro-écorégion. A terme nous envisageons d'élargir ces motifs aux pressions hydromorphologiques.

MOTS CLES: rivière, pressions physico-chimiques, indices biologiques, macro-invertébrés, poissons, diatomées, macrophytes, état écologique, Directive Cadre européenne sur l'Eau, pluridisciplinarité, fouille de données, motifs temporels.

ABSTRACT

Data from the assessment of river are big data, with complex relationships. Unsupervised data mining methods can be applied on them and give relevant results for their management, if a close collaboration exists between hydroecologists and computer scientists. The extraction of partially ordered patterns from temporal sequences of physicochemical pressures preceding a biological state has been achieved. These temporal patterns allow to identify a part of the pressures involved or not in a degraded ecological status, to specify the importance of the sequences time-length before a biological assessment, to identify the characteristic pressure categories at a regional scale. To go further, we plan to extend these patterns to hydromorphological pressures.

KEYWORDS: Rivers, physico-chemical pressures, biological indices, macroinvertebrates, fish, diatoms, macrophytes, ecological status, European Water Framework Directive, pluridisciplinarity, data mining, temporal patterns.

Université de Strasbourg

ENGEES

Image Ville Environnement

ED 413

Ecole Doctorale des Sciences de la Terre et de l'Environnement

Laboratoire Image Ville et Environnement LIVE UMR 7362

THÈSE

présentée par :

Corinne GRAC

Soutenue le :

19 décembre 2019

Pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline : Géochimie

Fouille temporelle des indicateurs physico-chimiques et biologiques pour l'évaluation de l'état, des pressions et de la capacité de résilience des rivières

ANNEXES

Rapporteurs

Philippe USSEGLIO-POLATERA	Professeur, Université de Lorraine, Metz
Yorick REYJOL	Chef d'équipe UMS Patrinat, MNHN, AFB, CNRS, Paris

Examinatrices

Gabrielle THIÉBAUT	Professeure, Université de Rennes
Maguelonne TEISSEIRE	Directrice de recherche, IRSTEA, Montpellier

Directeurs de thèse

Jean-Nicolas BEISEL	Professeur, ENGEES, Strasbourg
Michèle TRÉMOLIÈRES	Professeure émérite, Université de Strasbourg

Membre invitée

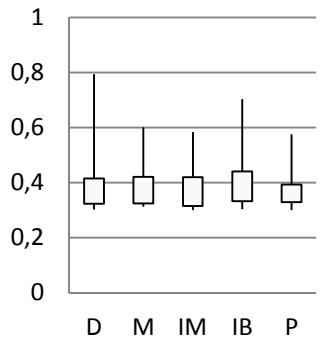
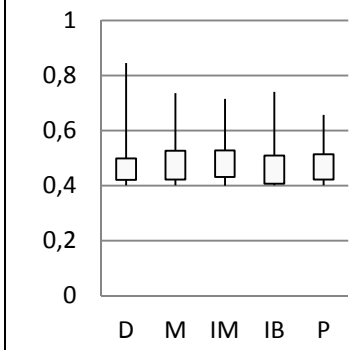
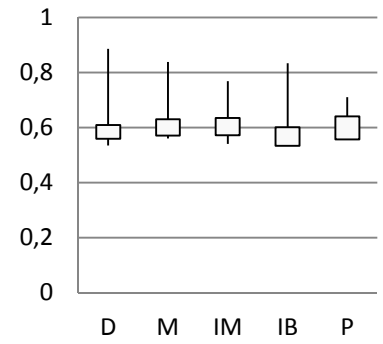
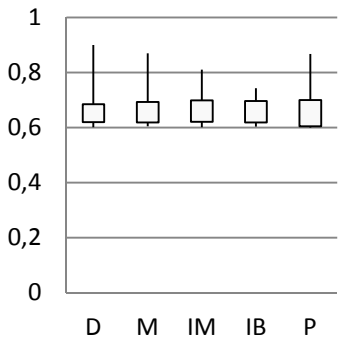
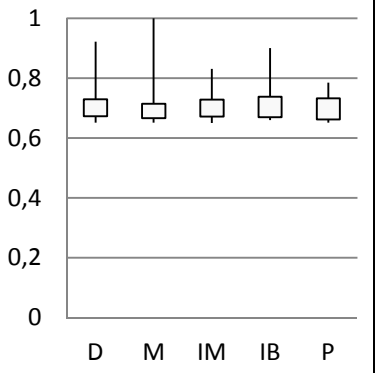
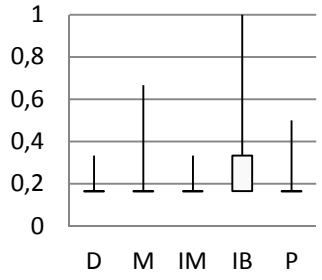
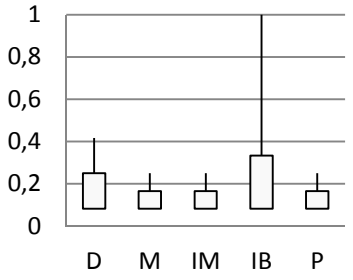
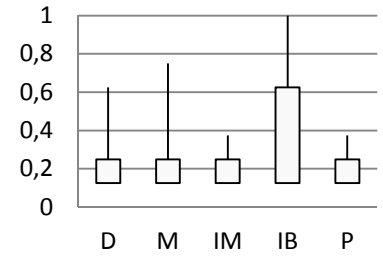
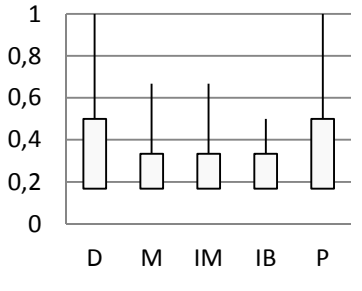
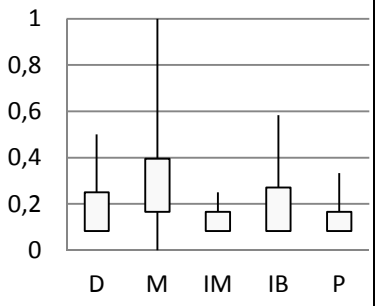
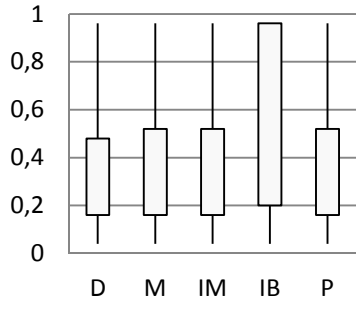
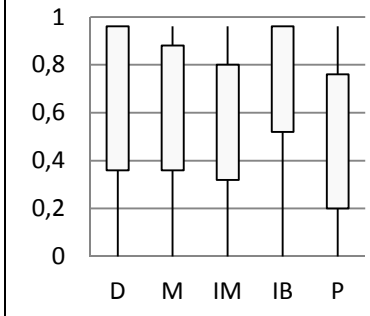
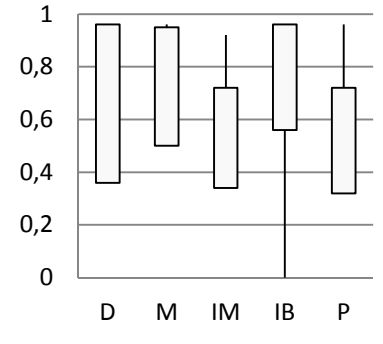
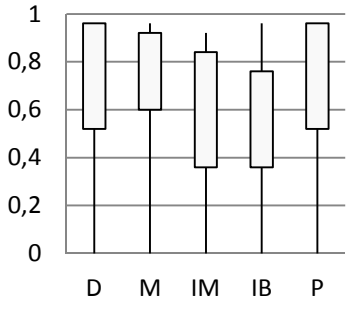
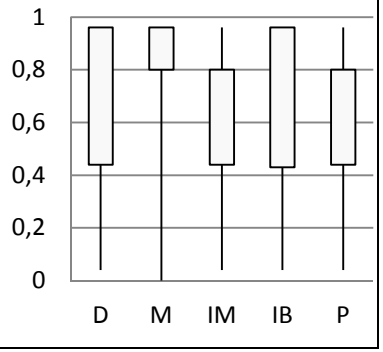
Florence LE BER	Ingénieure Générale des Ponts, des Eaux et des Forêts, ENGEES, Strasbourg

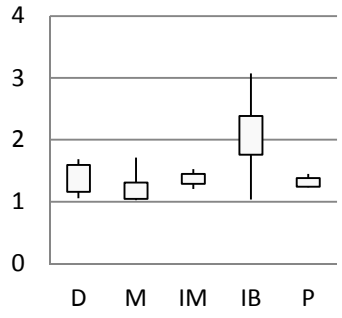
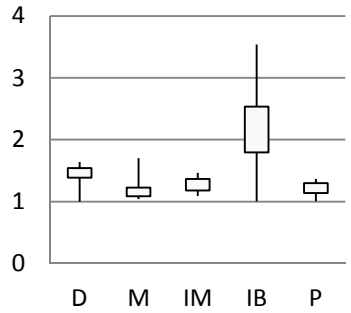
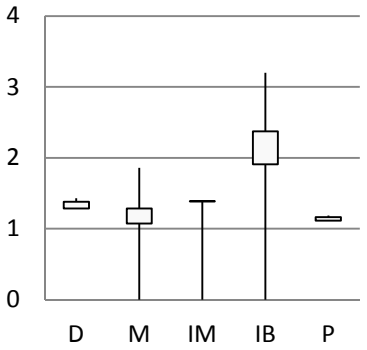
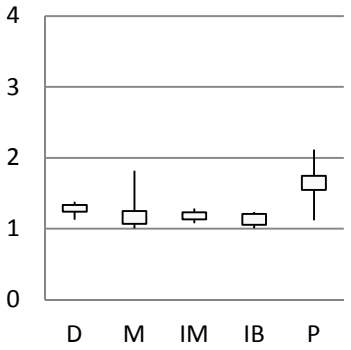
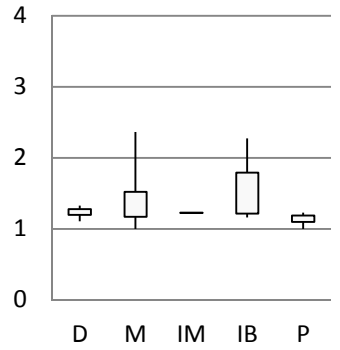
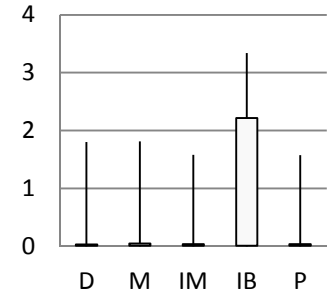
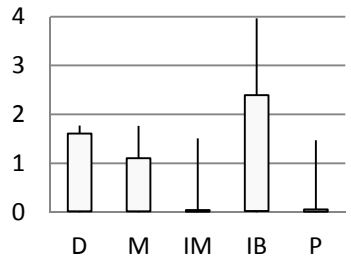
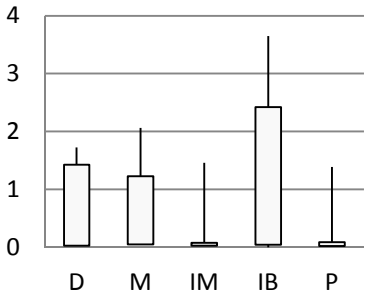
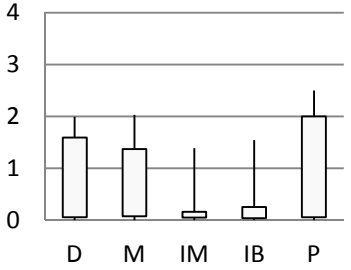
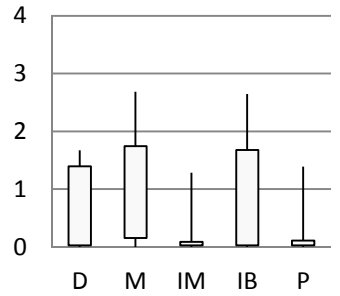
SOMMAIRE

LISTE DES SIGLES & ACRONYMES

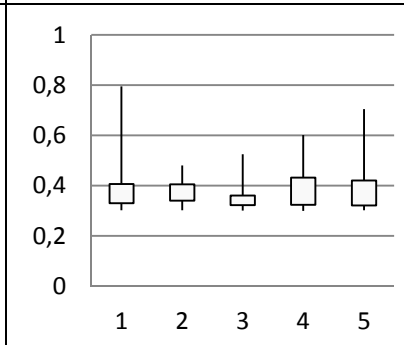
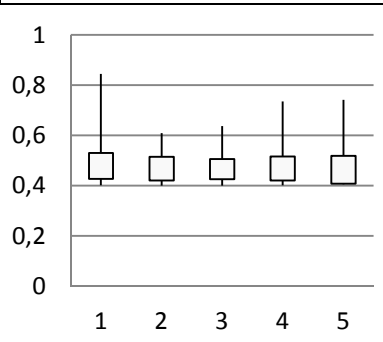
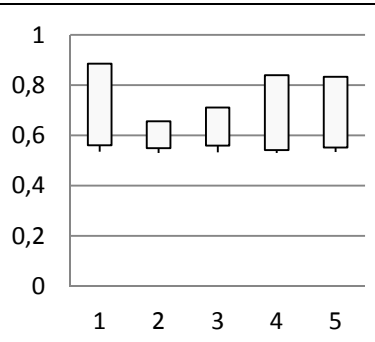
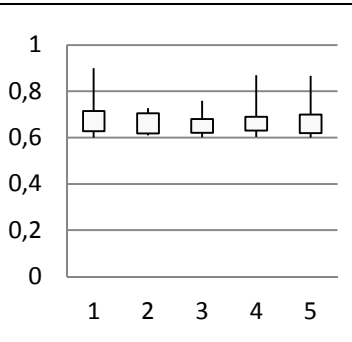
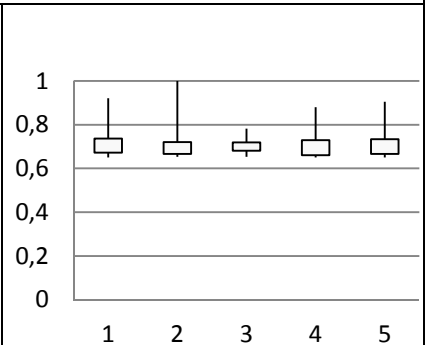
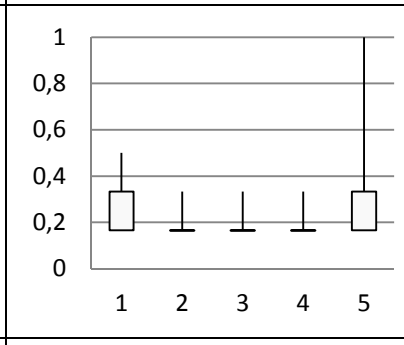
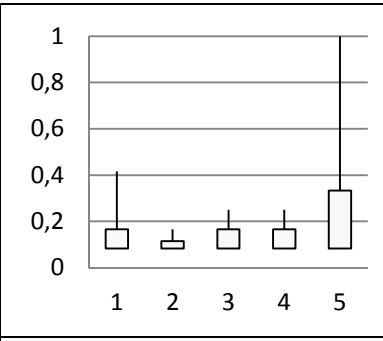
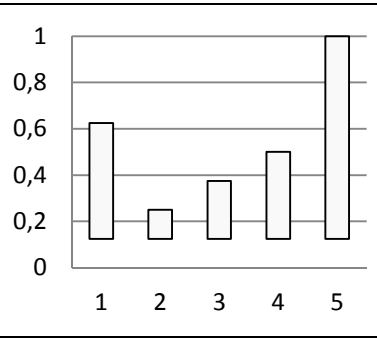
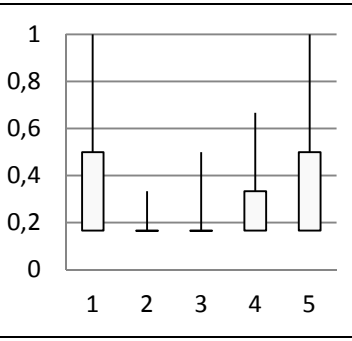
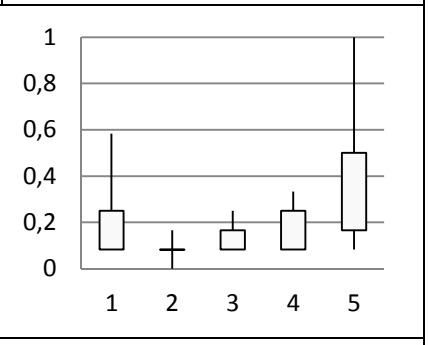
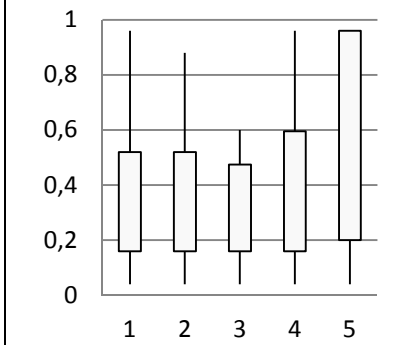
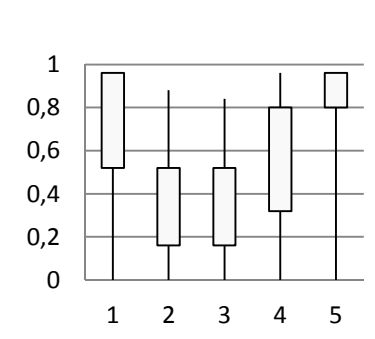
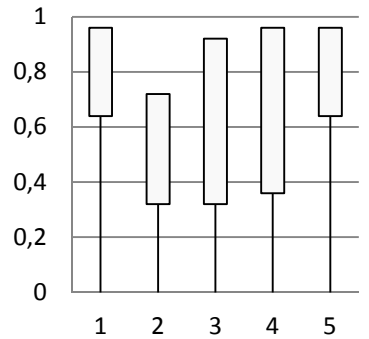
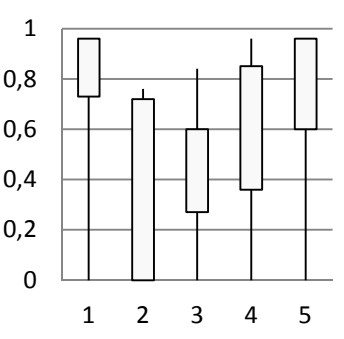
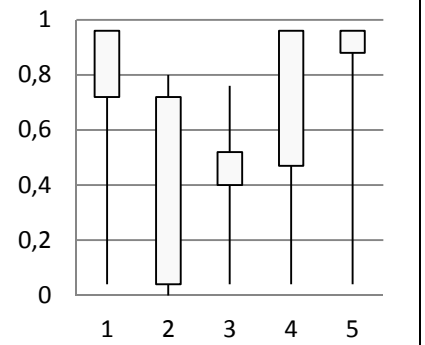
ACID : Acidification, altération du SEQ-eau et des grilles DCE
AZOT : Matières azotées hors nitrates, altération du SEQ-eau
BILO2 : Bilan oxygène, altération des grilles DCE
EPRV : Effets Prolifération Végétale, altération du SEQ-eau
HAP : Hydrocarbures aromatiques polycycliques, altération du SEQ-eau
HER: Hydro-écorégion
I2M2 : Indice Invertébrés Multi-Métrique
IBD: Indice Biologique Diatomique
IBGN: Indice Biologique Global Normalisé
IBMR : Indice Biologique Macrophytique en Rivière
IPR: indice Poissons en Rivière
MINE : Paramètres de la minéralisation, altération du SEQ-eau
MOOX : Matières organiques et oxydables, altération du SEQ-eau
MPMI : Micropolluants minéraux, altération du SEQ-eau
MPOR : Micropolluants organiques hors pesticides, altération du SEQ-eau
NITR: Nitrates, altération du SEQ-eau
NUTRI : Nutriments, altération des grilles DCE
PAES : Particules en suspension, altération du SEQ-eau
PCB : Poly-chloro-biphényles, altération du SEQ-eau
PEST : Pesticides, altération du SEQ-eau
PHOS : Matières phosphorées, altération du SEQ-eau
POSPE : Polluants Spécifiques, altération des grilles DCE
PRESTOR: temporal PRESsure categories patterns extracTOR
SDP : Substances prioritaires et substances dangereuses prioritaires, altération des grilles DCE
SEQ-eau : Système d'Evaluation de la qualité de l'eau
TEMP : Température, altération du SEQ-eau et des grilles DCE

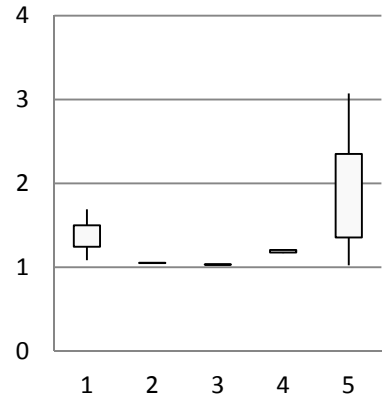
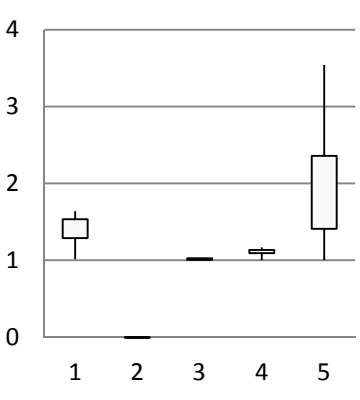
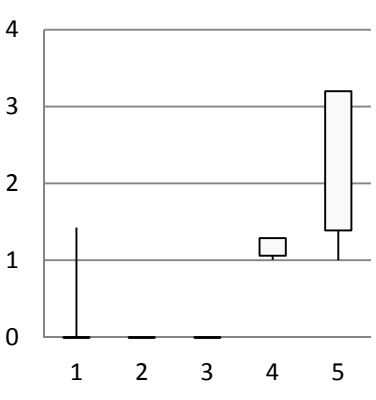
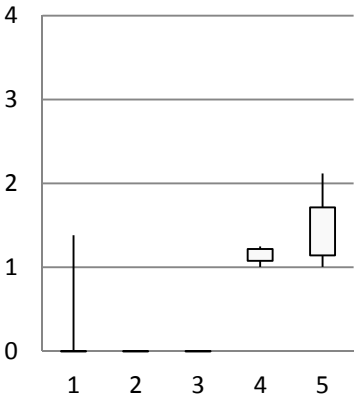
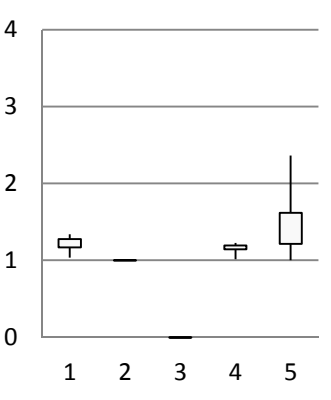
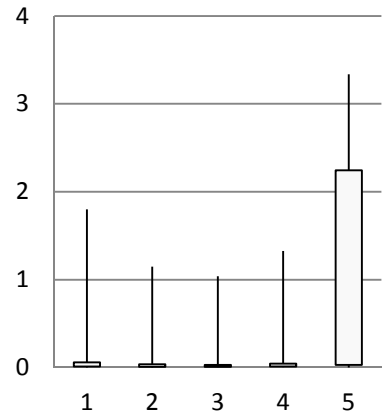
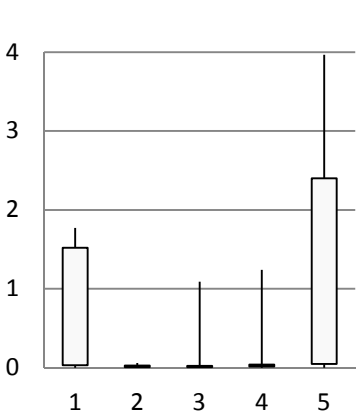
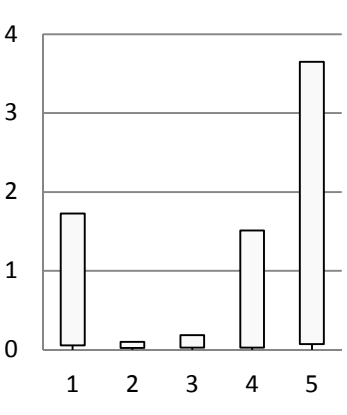
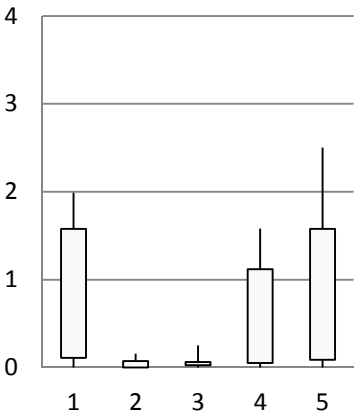
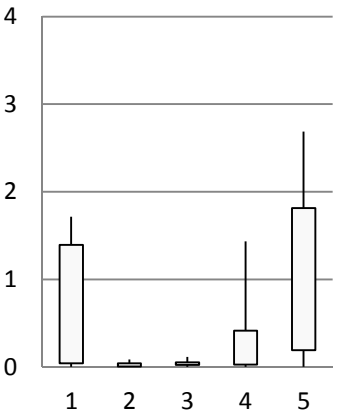
ANNEXE 1 : Variations des mesures d'intérêt obtenues pour les extractions faites pour le SEQ-eau, pour les longueurs de séquences 3, 6, 12, 18 et 24 mois par indices biologiques (D=IBD, M=IBMR, IM= I2M2, IB=IBGN, P=IPR)

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
F					
C					
S					

	3 MOIS	6 MOIS	12 MOIS	18 MOIS	24 MOIS
E					
FCS+E					

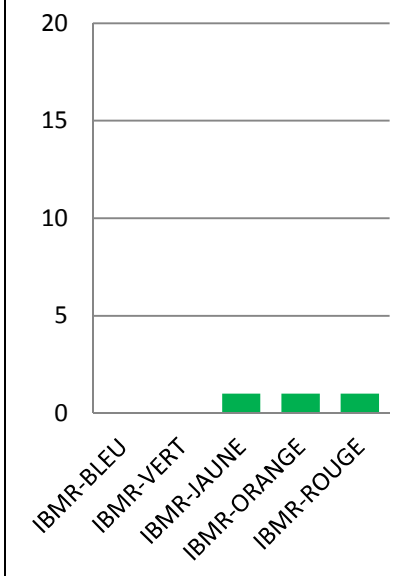
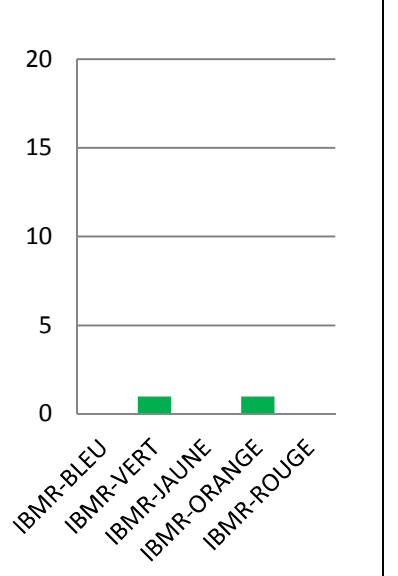
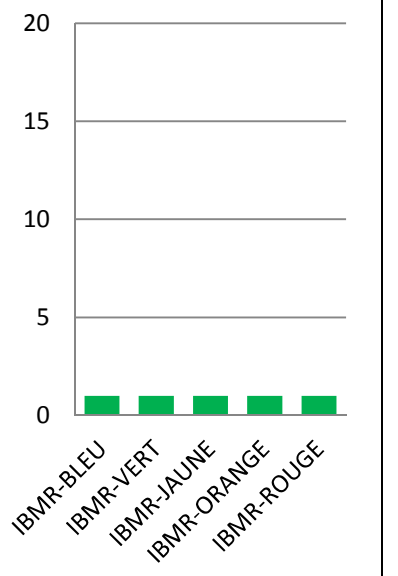
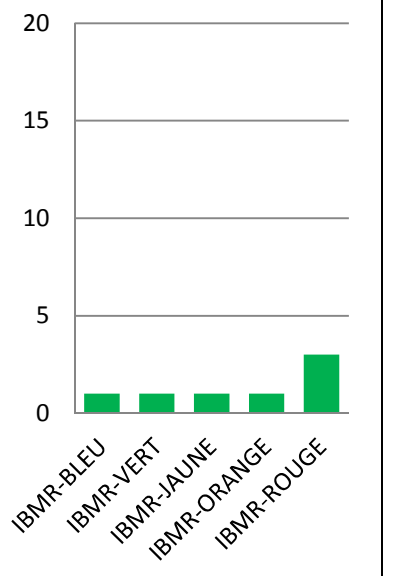
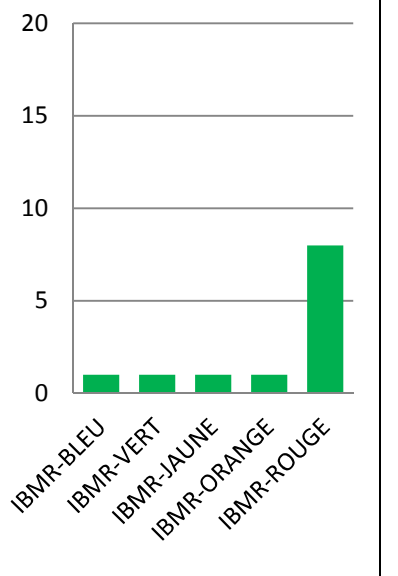
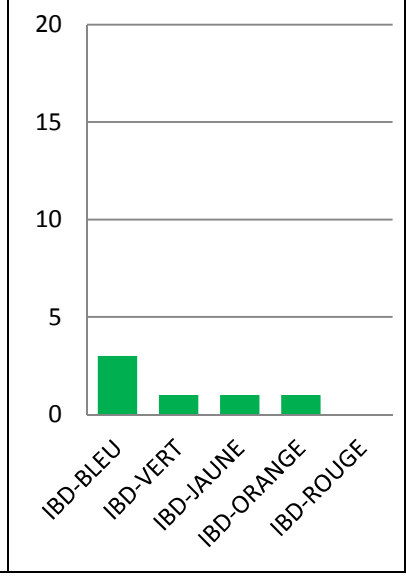
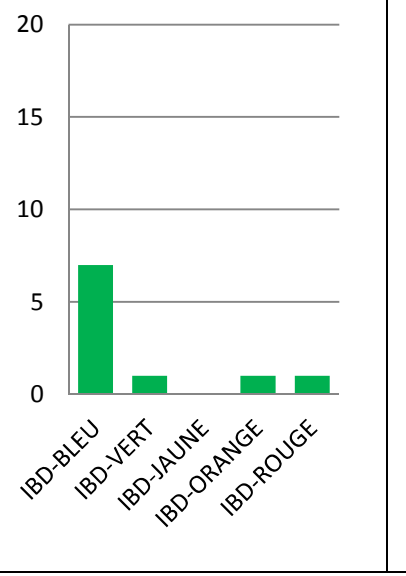
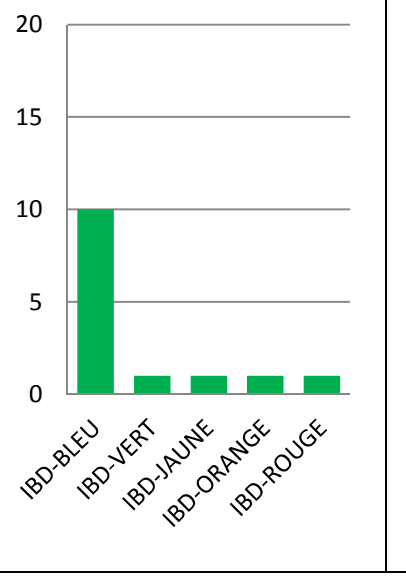
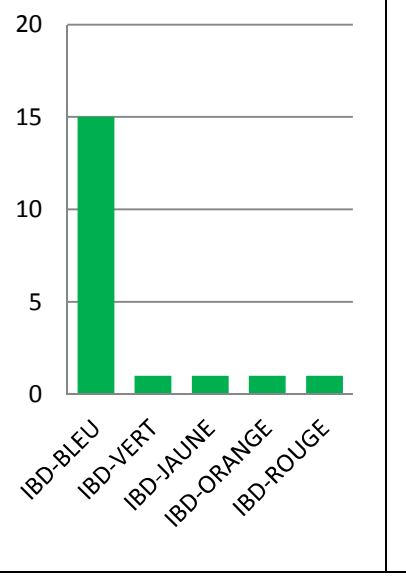
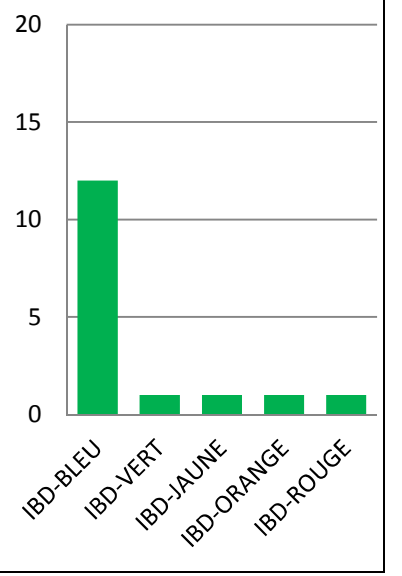
ANNEXE 2 : Variations des mesures d'intérêt obtenues pour les extractions faites pour le SEQ-eau, pour les longueurs de séquences 3, 6, 12, 18 et 24 mois par état biologique (de classe 1 : très bonne à classe 5 : mauvaise)

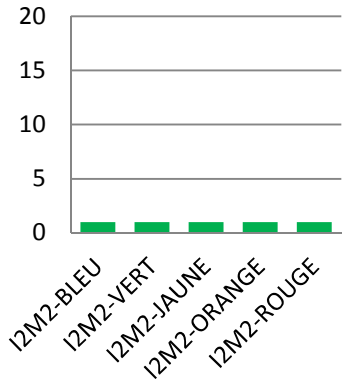
	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
F					
C					
S					

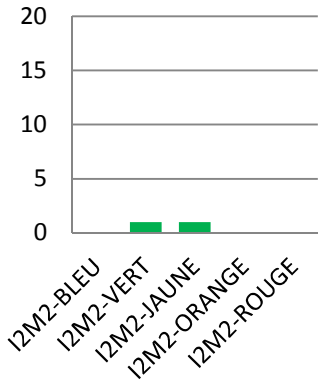
	3 MOIS	6 MOIS	12 MOIS	18 MOIS	24 MOIS
E					
FCS+E					

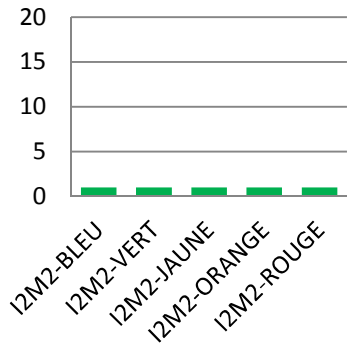
ANNEXE 3 : Nombre d'apparitions des items par altération, par indice biologique et par longueur de séquences 3, 6, 12, 18 et 24 mois, dans l'ensemble des motifs extraits pour le SEQ-eau

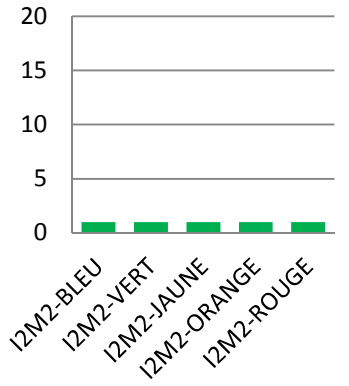
Altération ACID pour l'IBMR & l'IBD

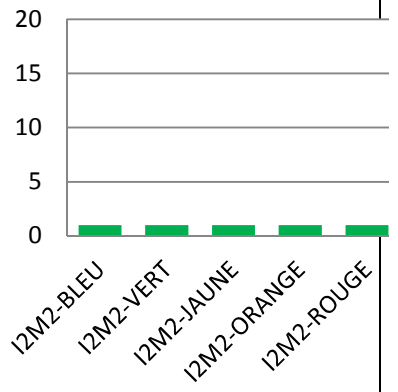
	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
IBMR					
IBD					

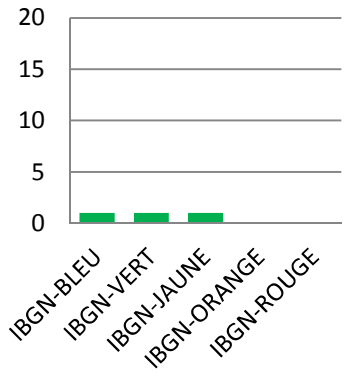


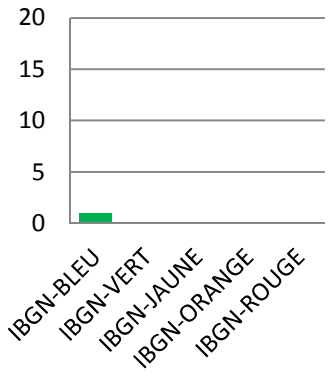


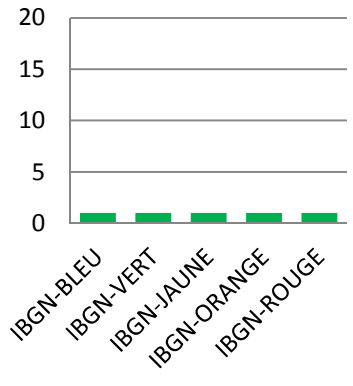


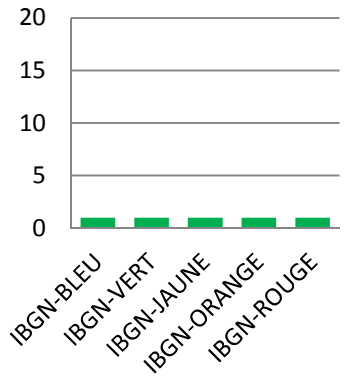


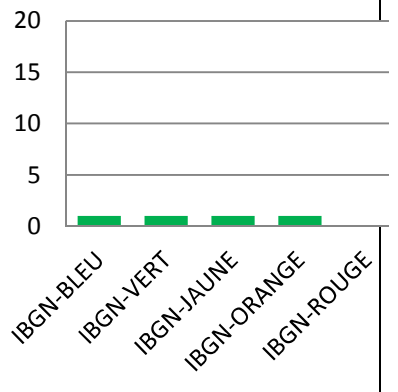


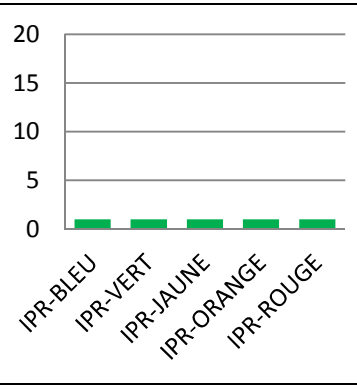


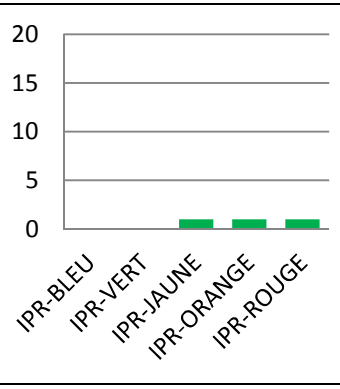


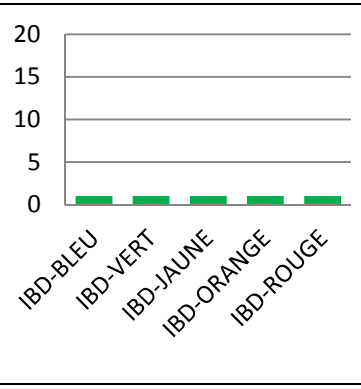


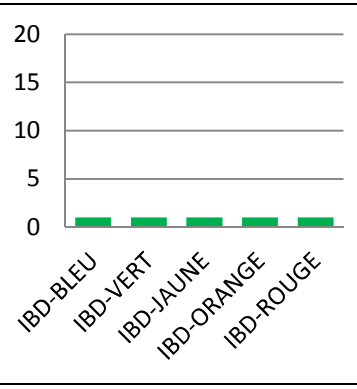


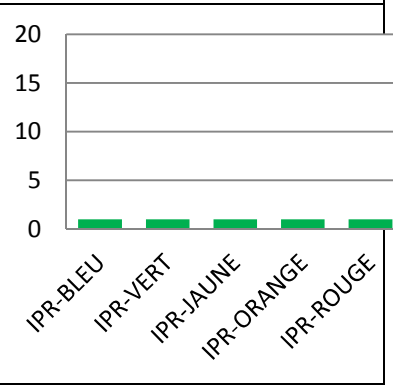






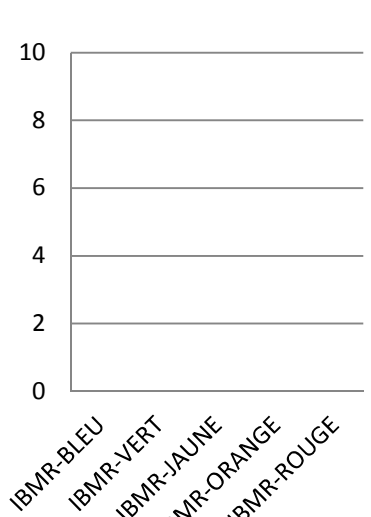
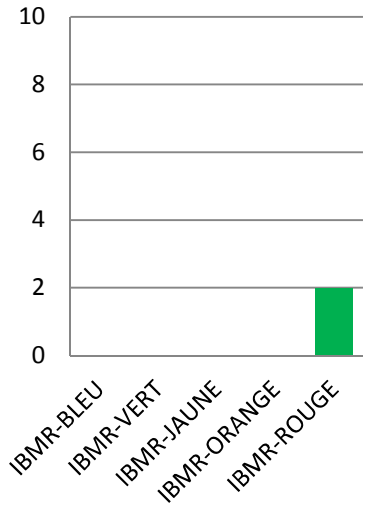
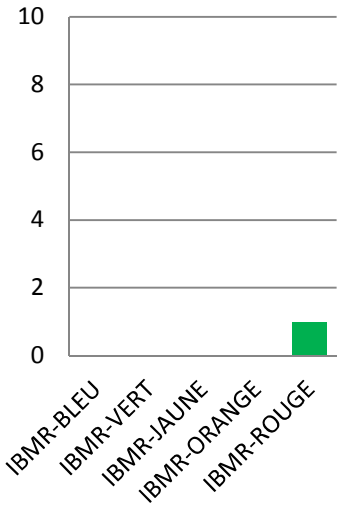
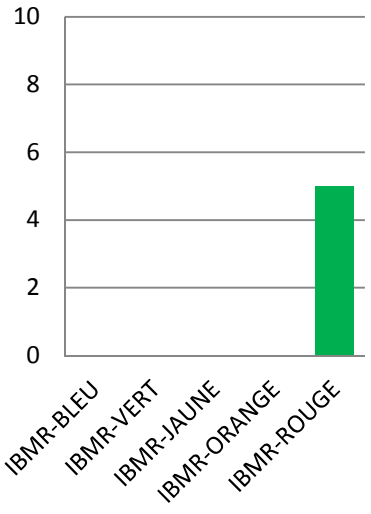
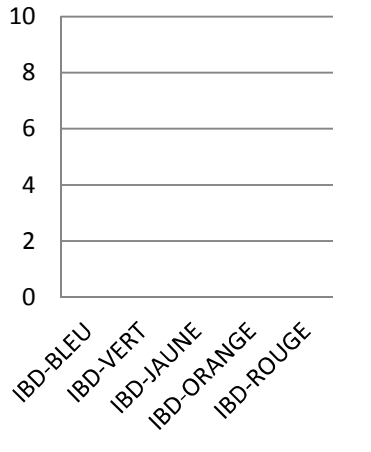
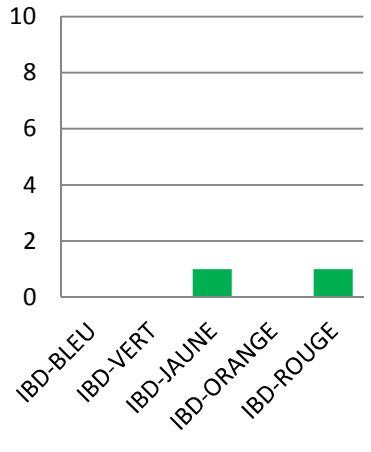
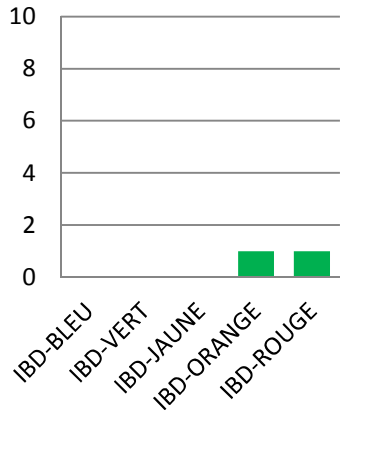
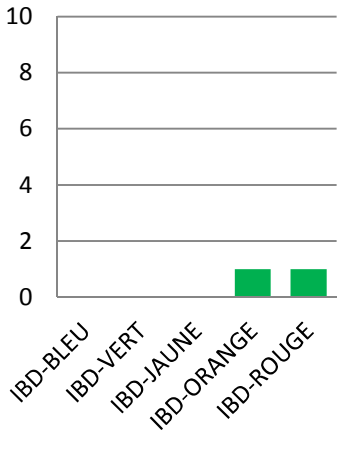
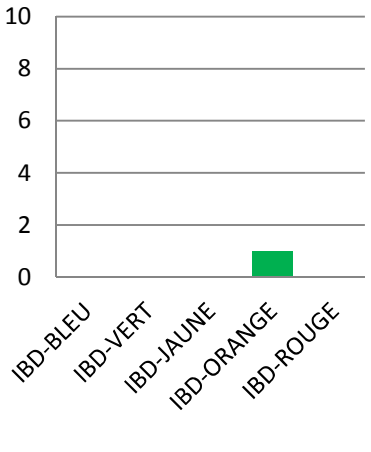




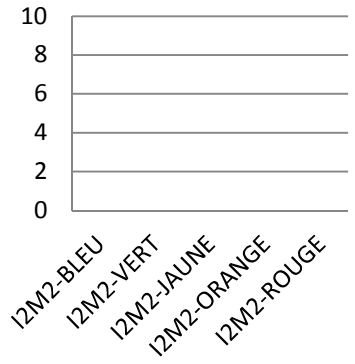
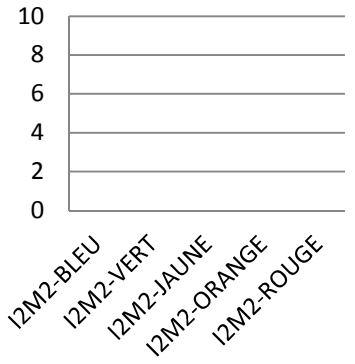
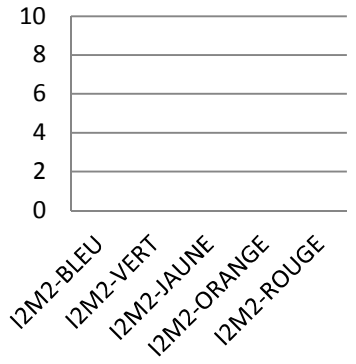
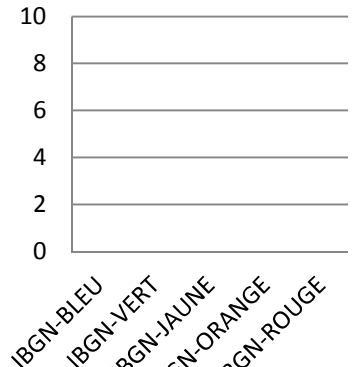
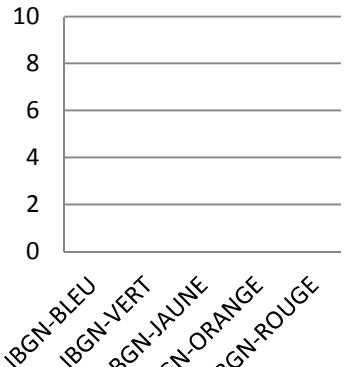
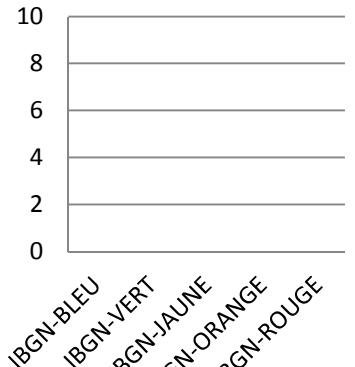
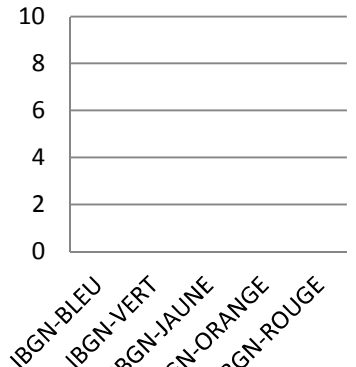
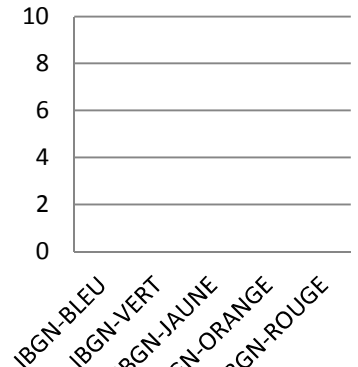
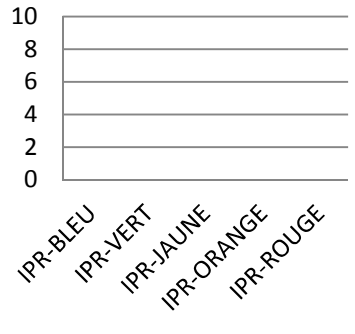
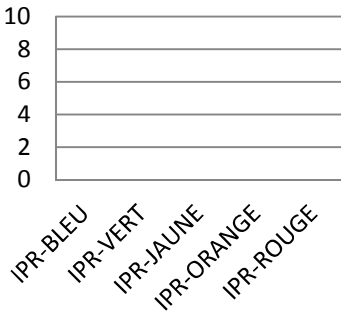
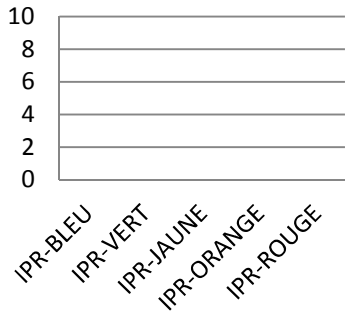
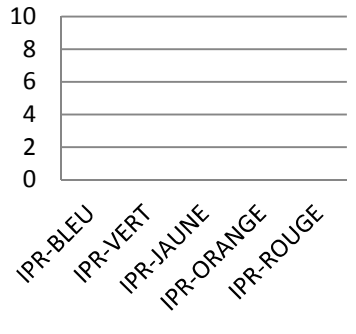
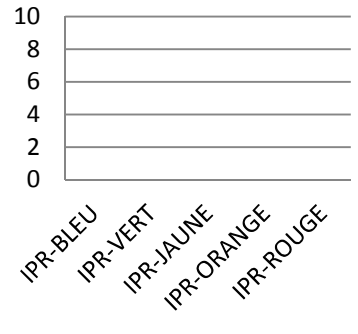


Altération **ACID** pour l'I2M2, l'IBGN & l'IPR

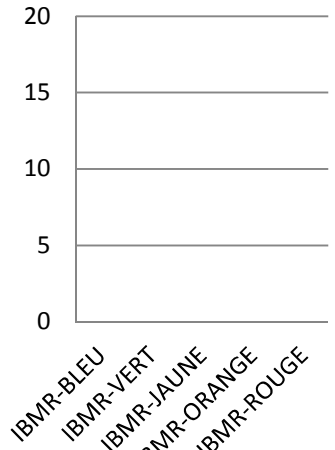
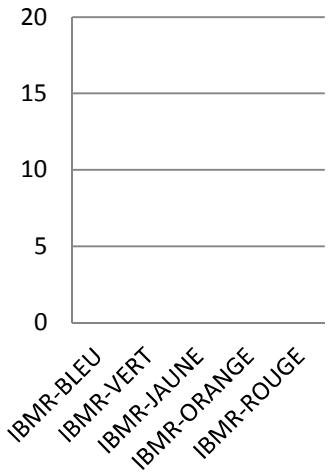
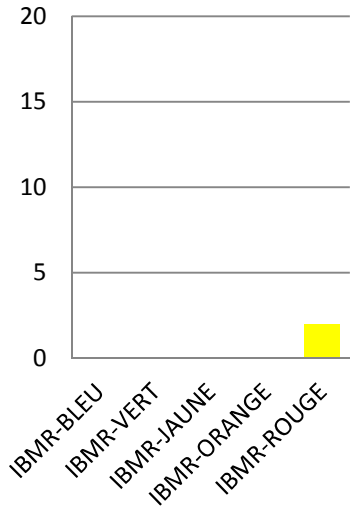
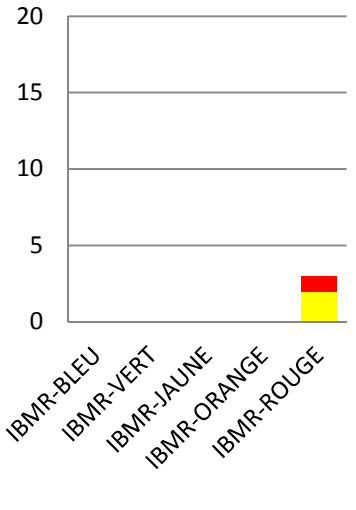
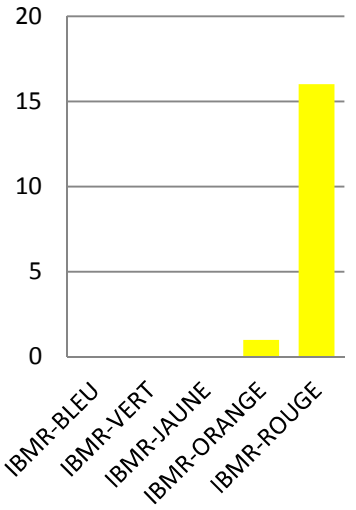
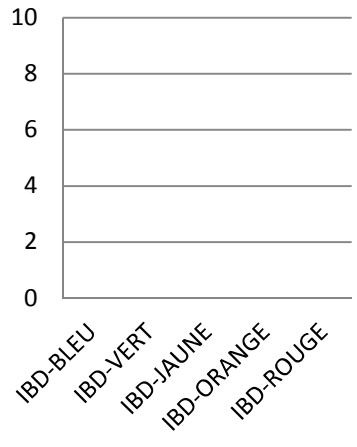
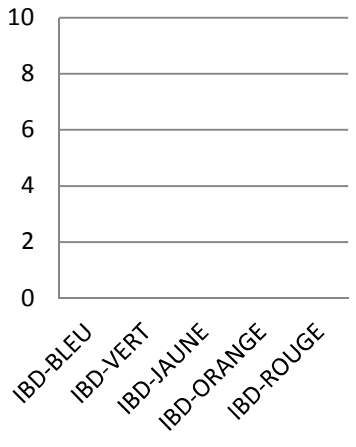
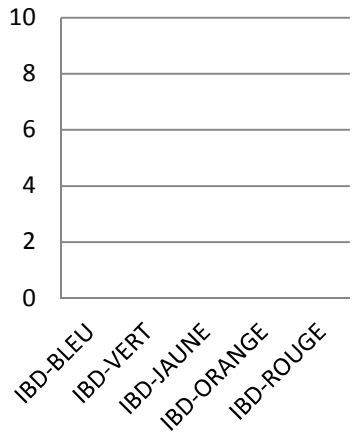
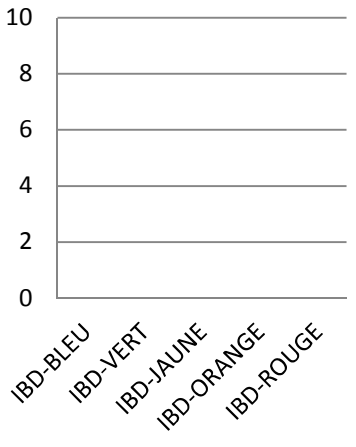
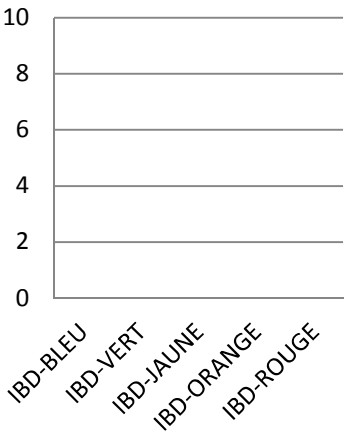
Altération MINE pour l'IBMR & l'IBD

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
IBMR					
IBD					

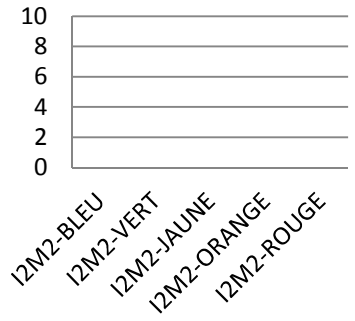
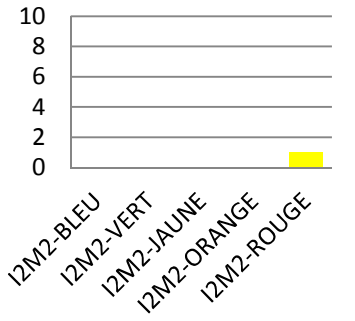
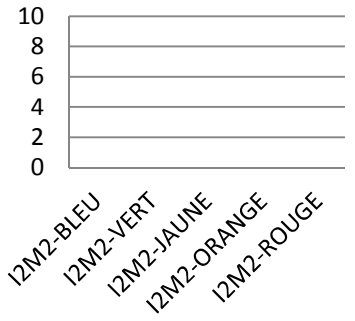
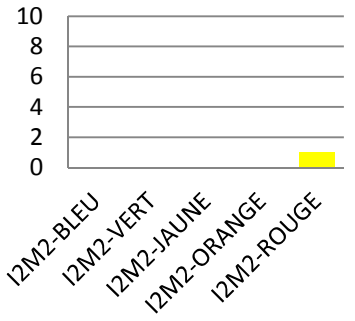
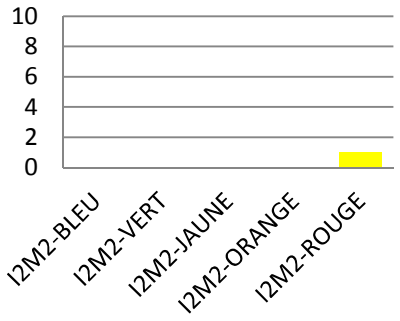
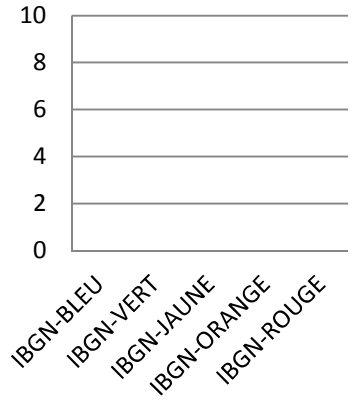
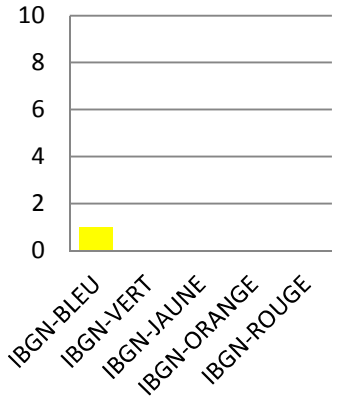
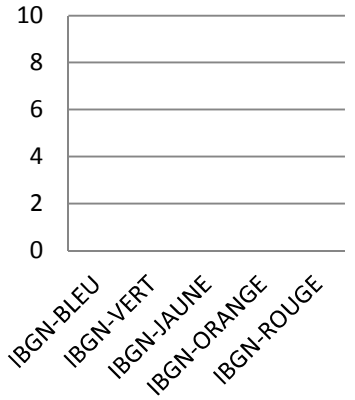
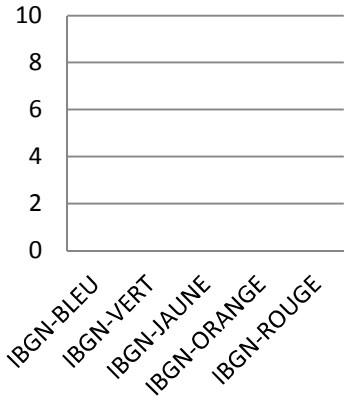
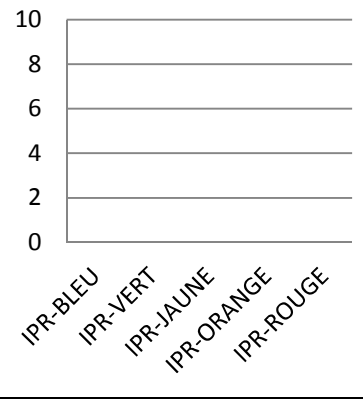
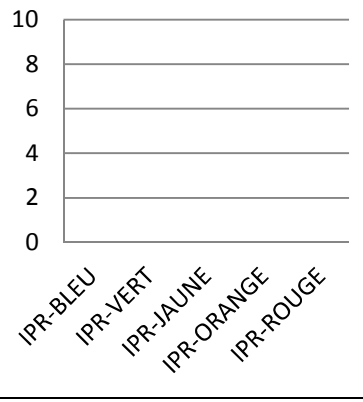
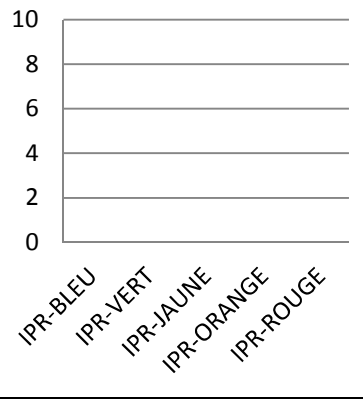
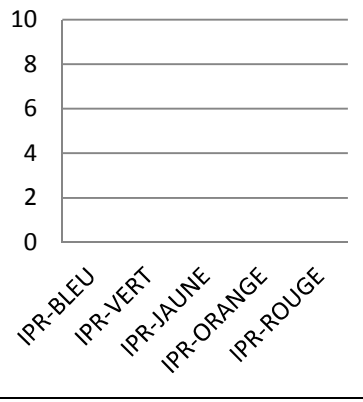
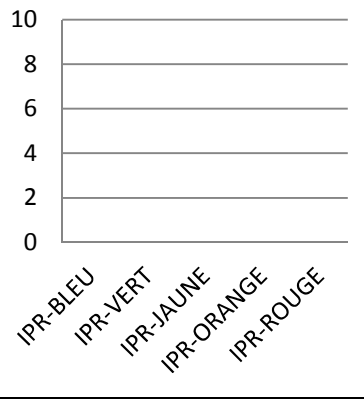
Altération **MINE** pour l'I2M2, l'IBGN & l'IPR

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
I2M2					
IBGN					
IPR					

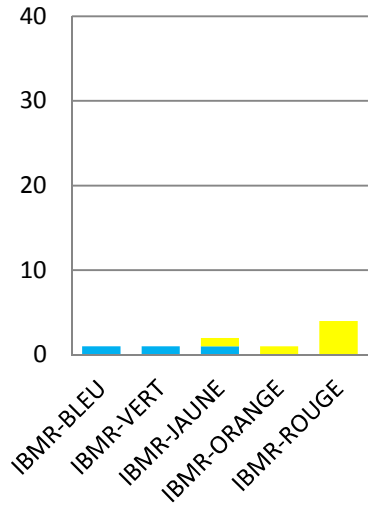
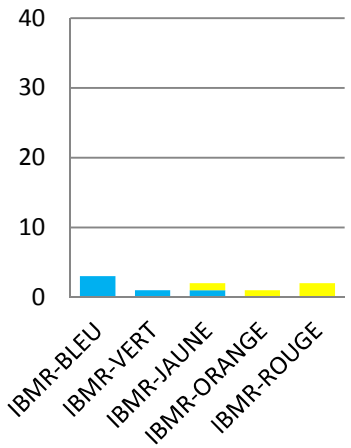
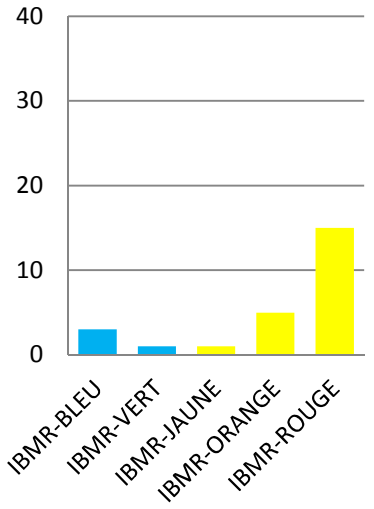
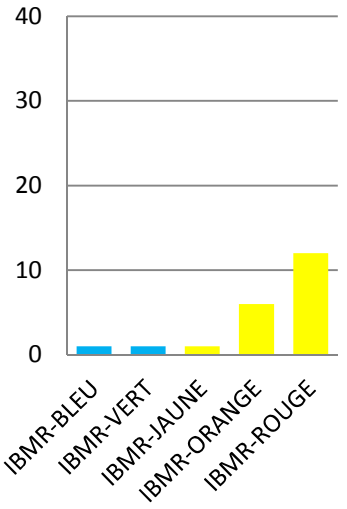
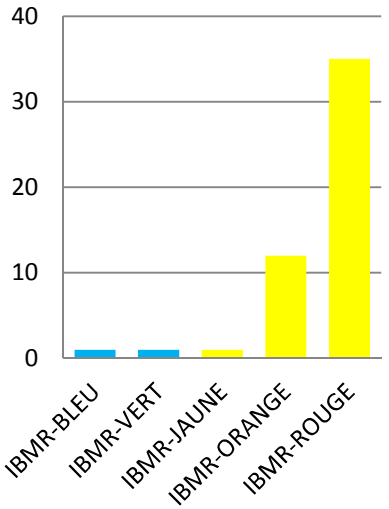
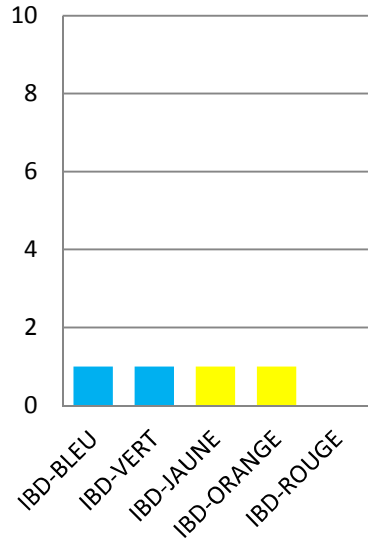
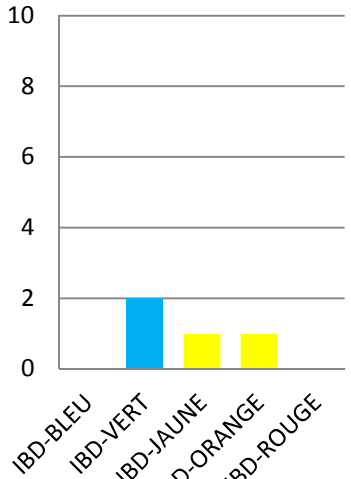
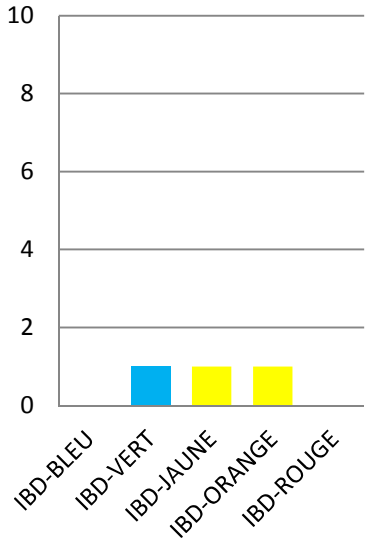
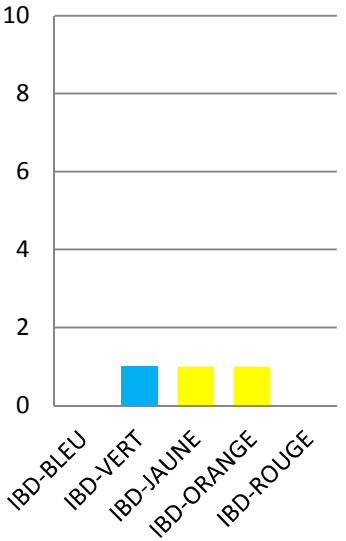
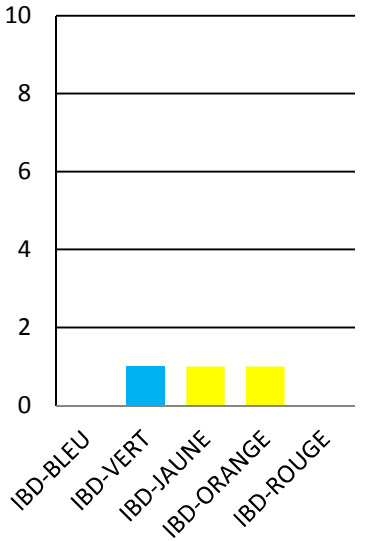
Altération PAES pour l'IBMR & l'IBD

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
IBMR					
IBD					

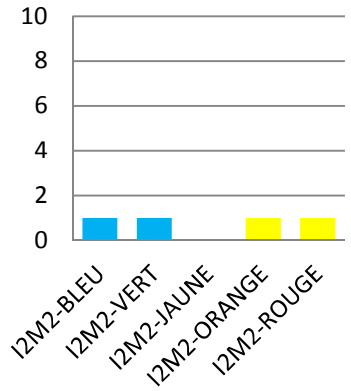
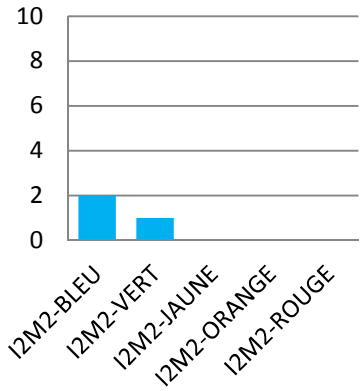
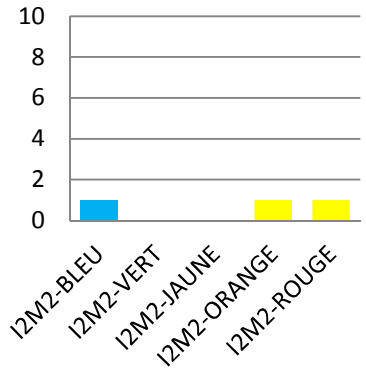
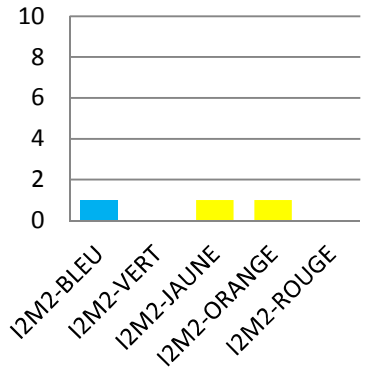
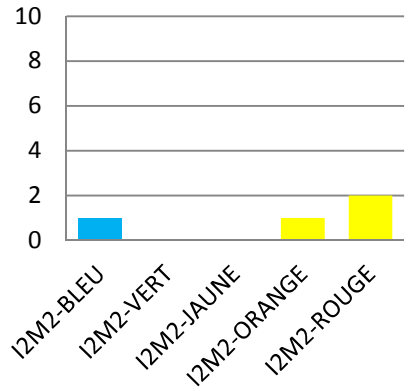
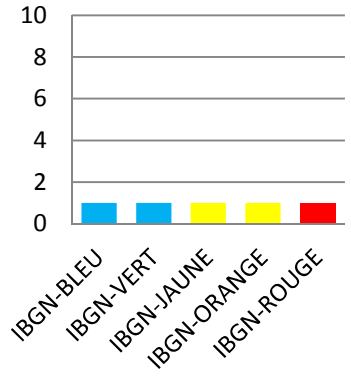
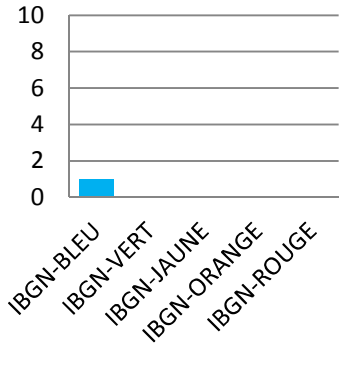
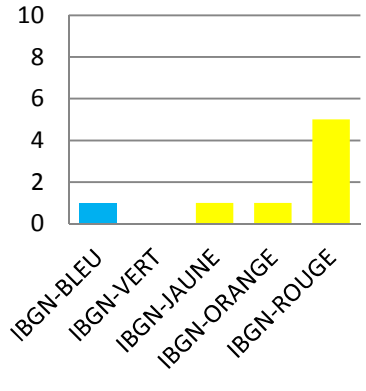
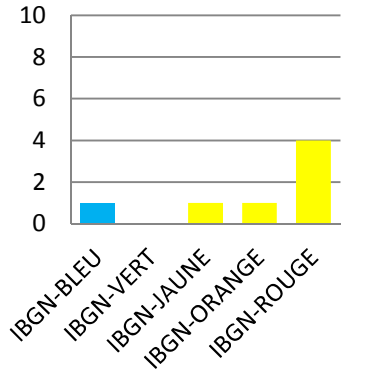
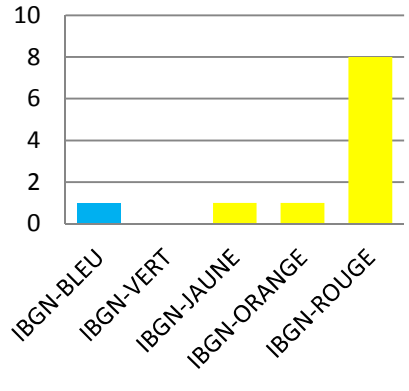
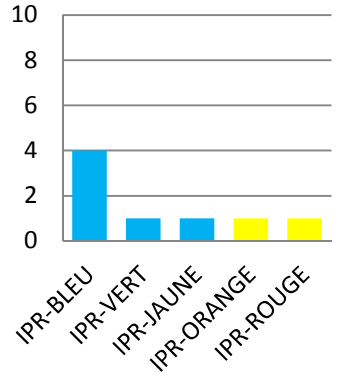
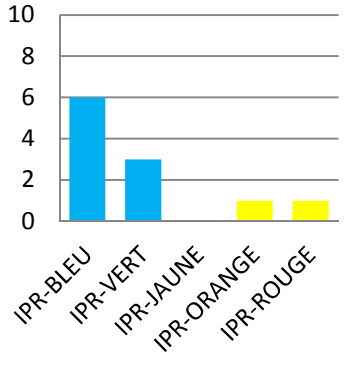
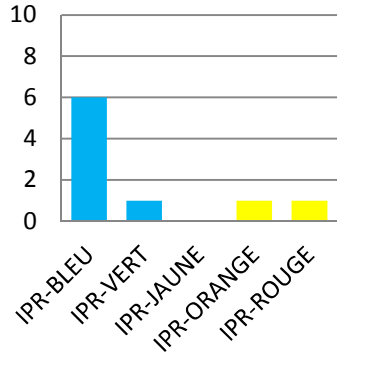
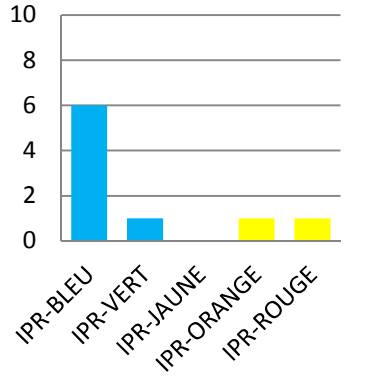
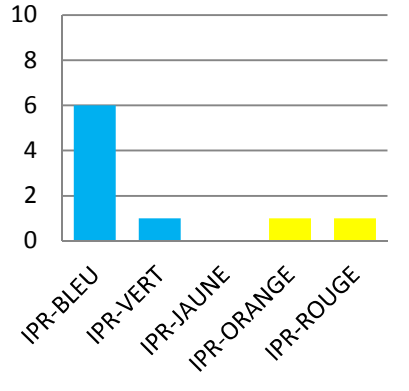
Altération PAES pour l'I2M2, l'IBGN & l'IPR

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
I2M2					
IBGN					
IPR					

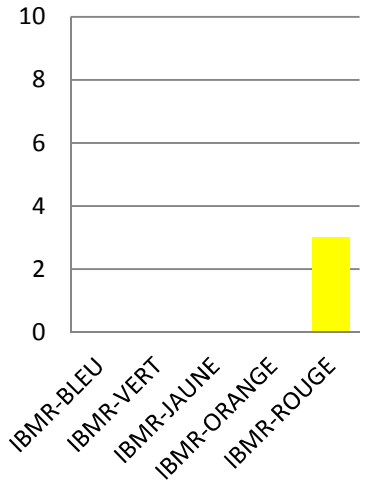
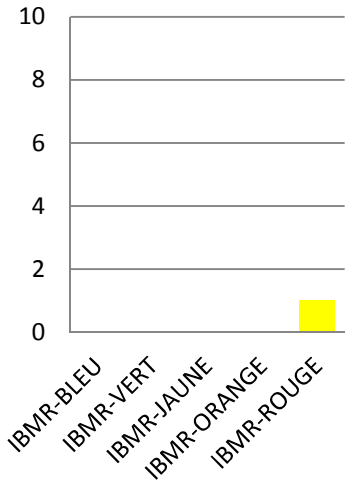
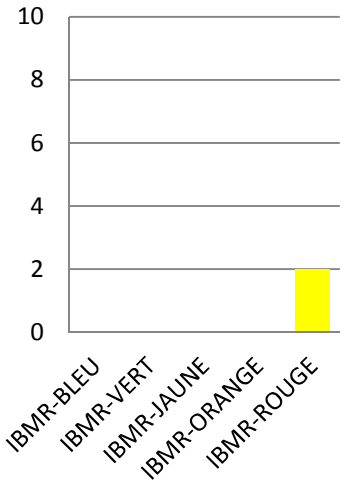
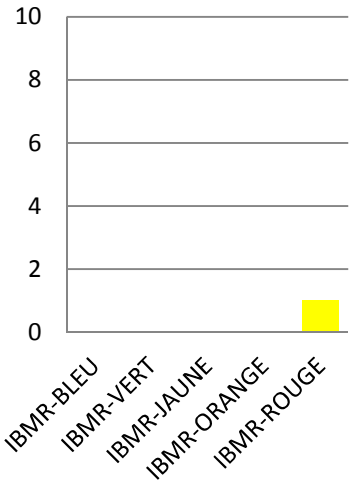
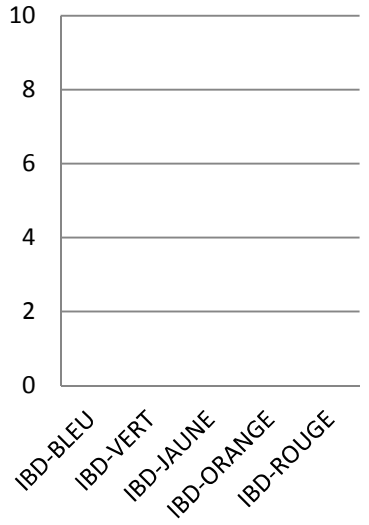
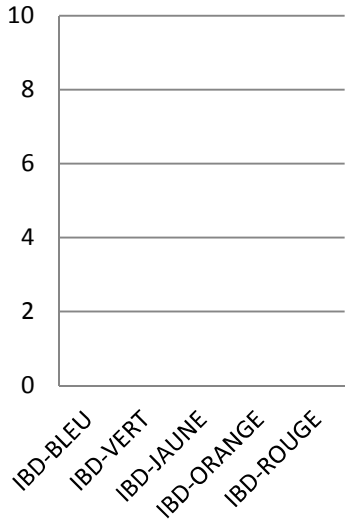
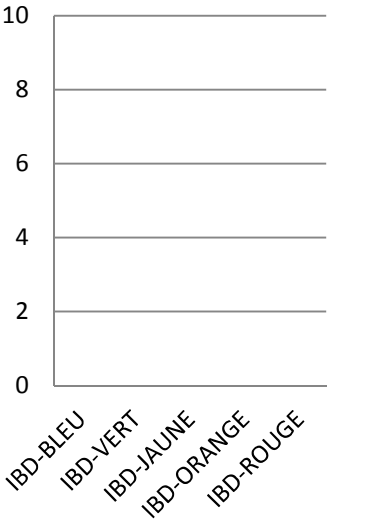
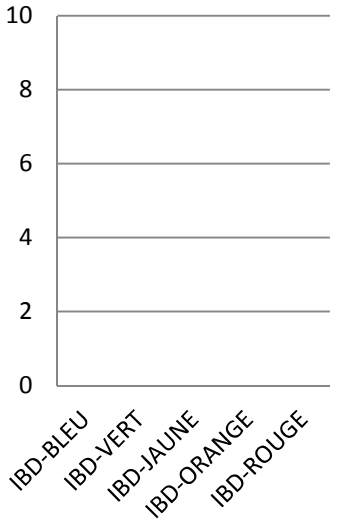
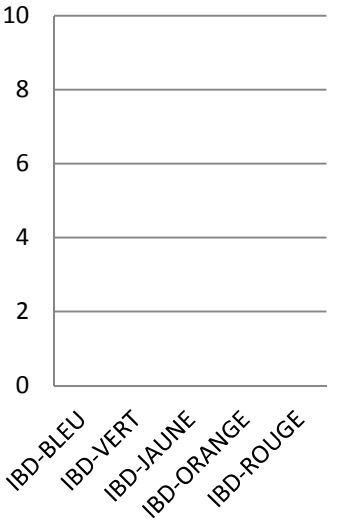
Altération MOOX pour l'IBMR & l'IBD

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
IBMR					
IBD					

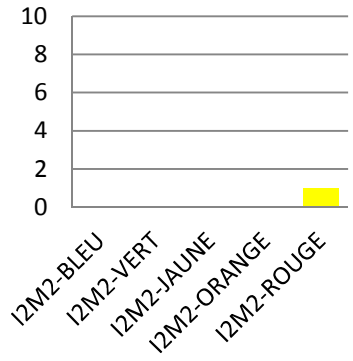
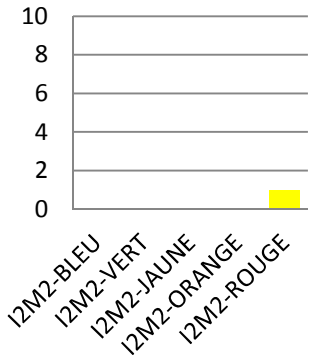
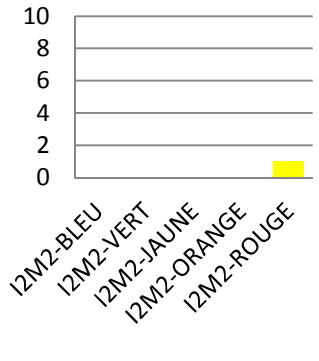
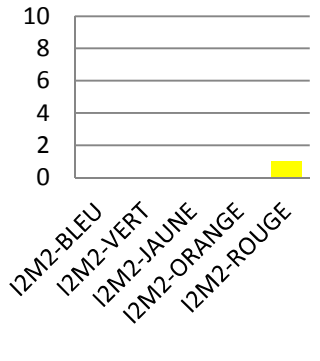
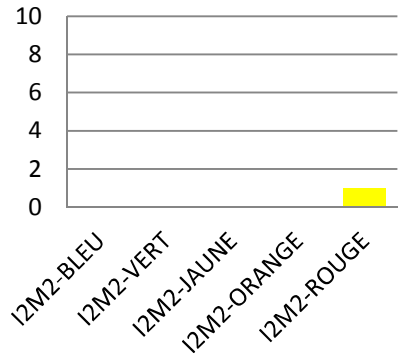
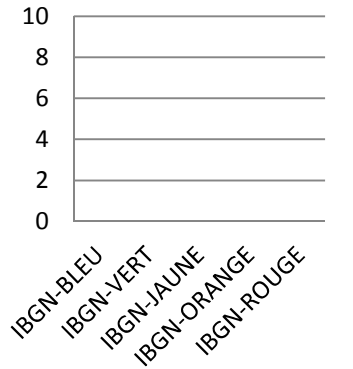
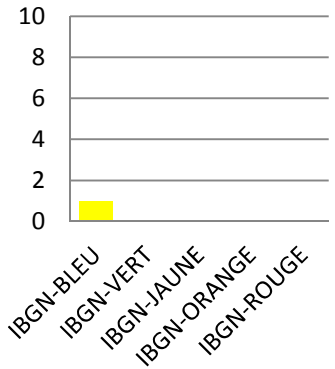
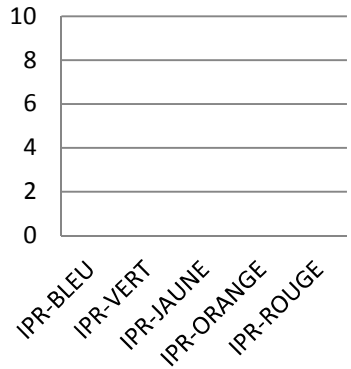
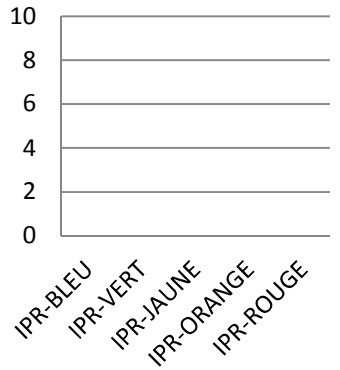
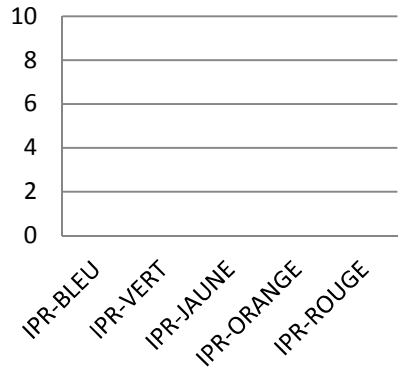
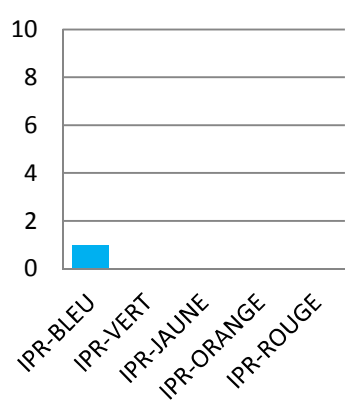
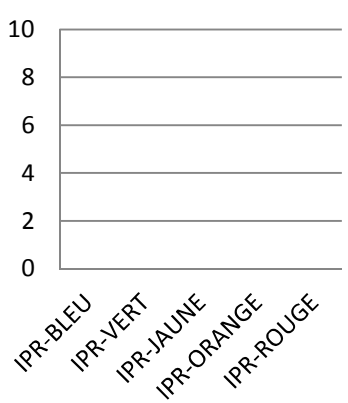
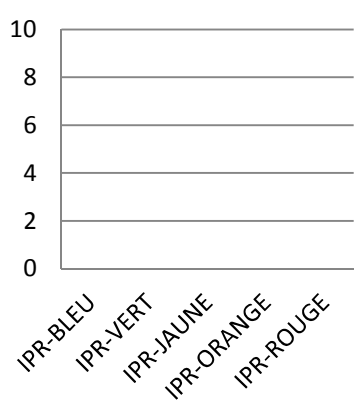
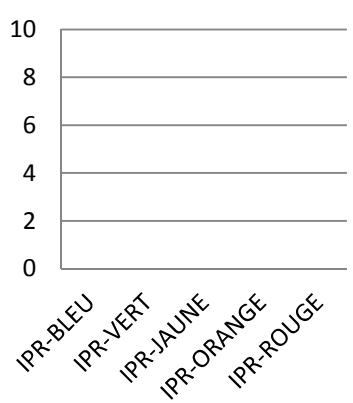
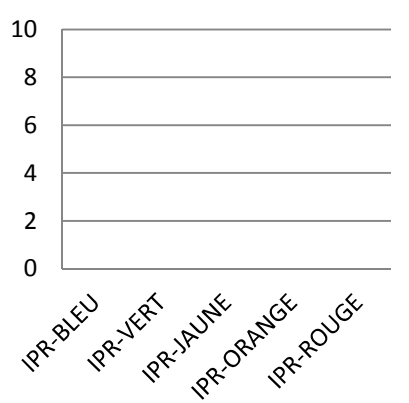
Altération **MOOX** pour l'I2M2, l'IBGN & l'IPR

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
I2M2					
IBGN					
IPR					

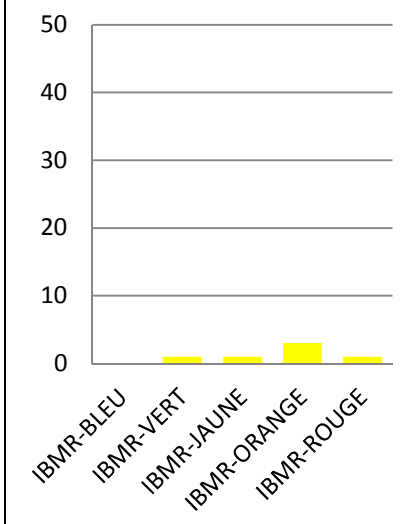
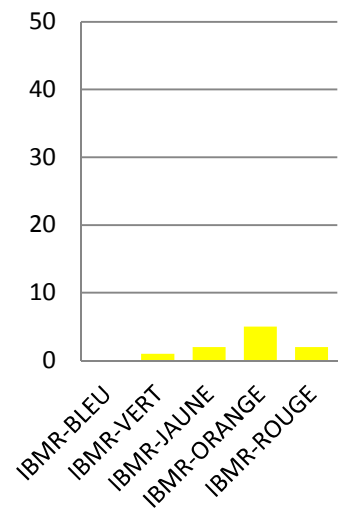
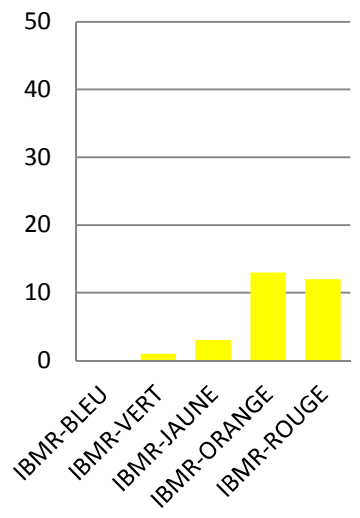
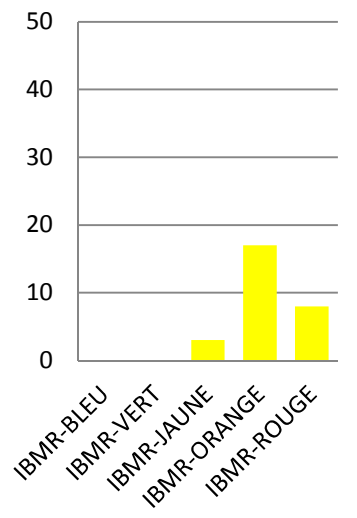
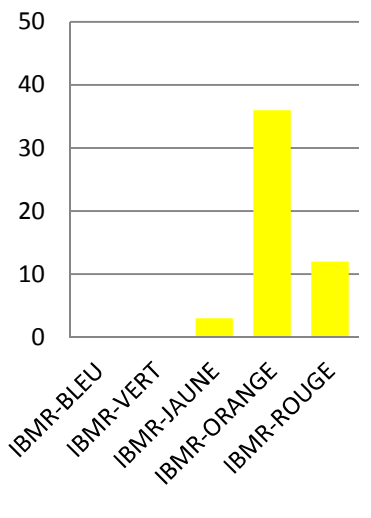
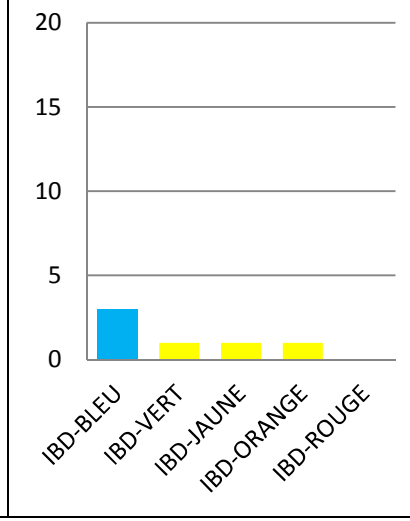
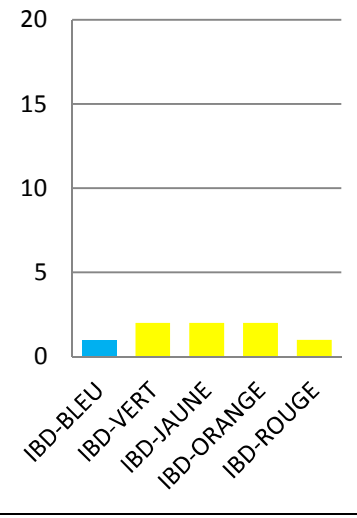
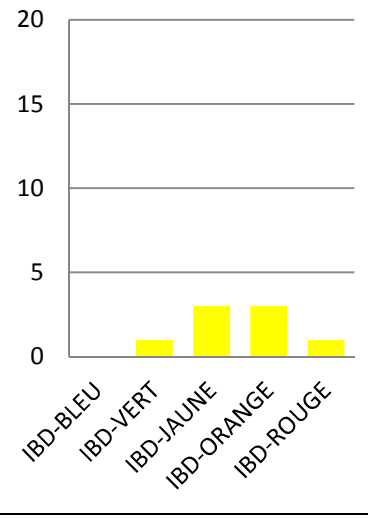
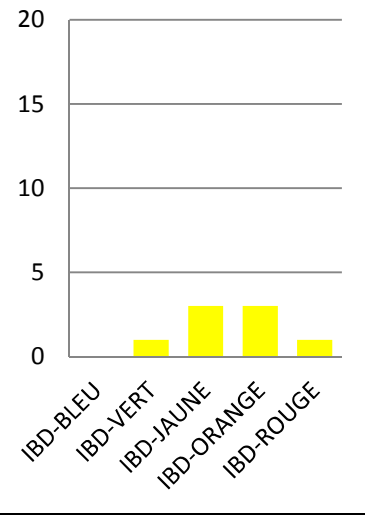
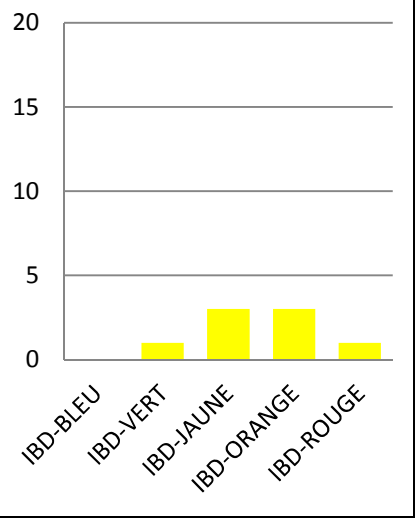
Altération AZOT pour l'IBMR & l'IBD

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
IBMR					
IBD					

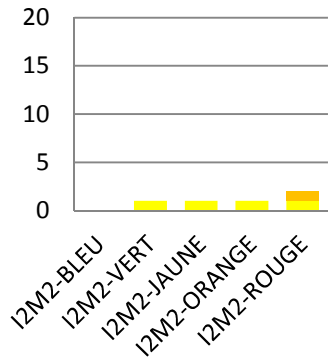
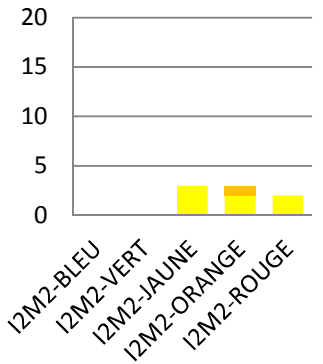
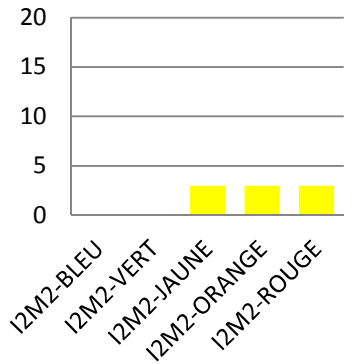
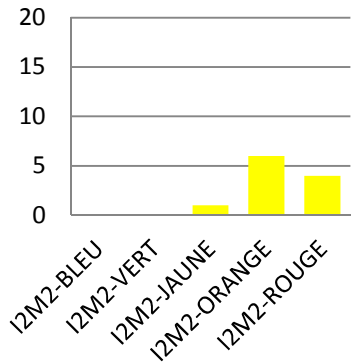
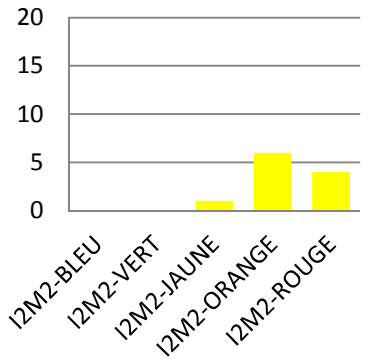
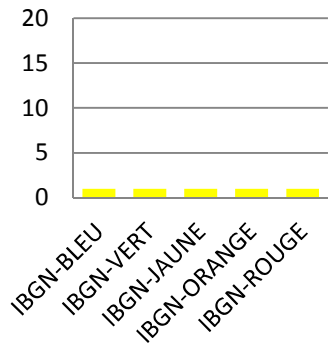
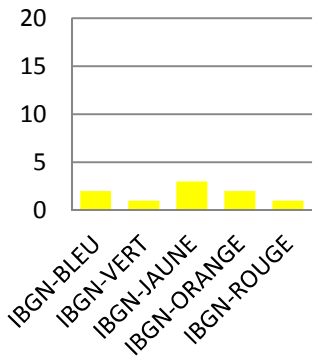
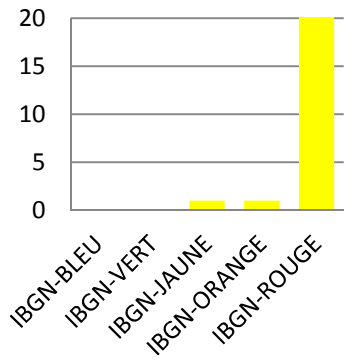
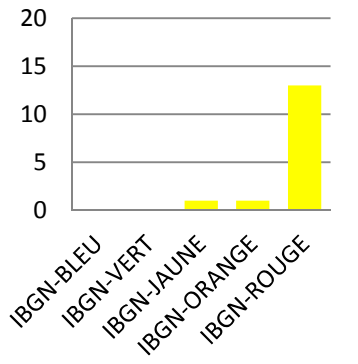
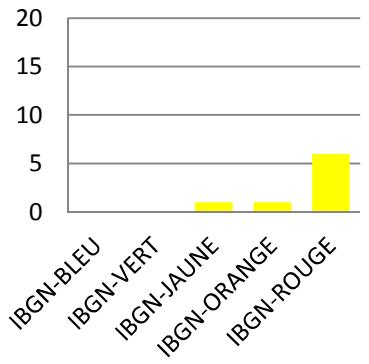
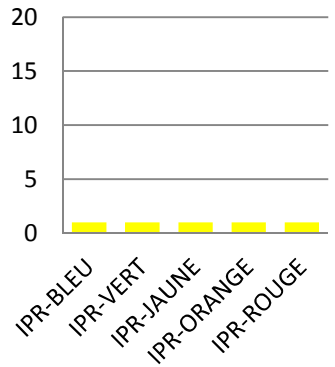
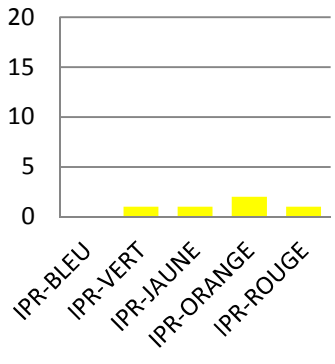
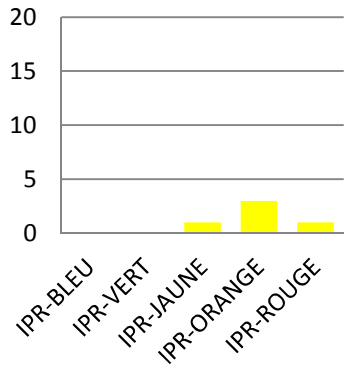
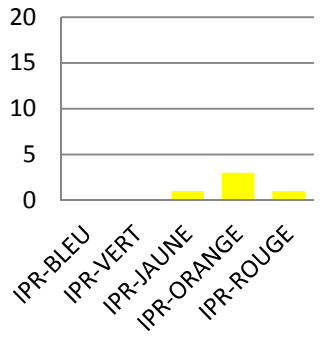
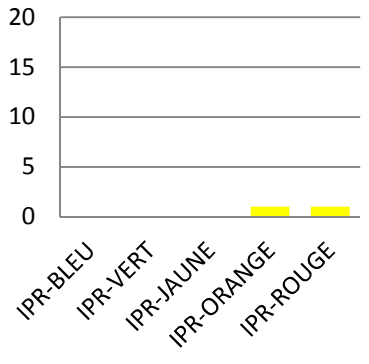
Altération **AZOT** pour l'I2M2, l'IBGN & l'IPR

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
I2M2					
IBGN					
IPR					

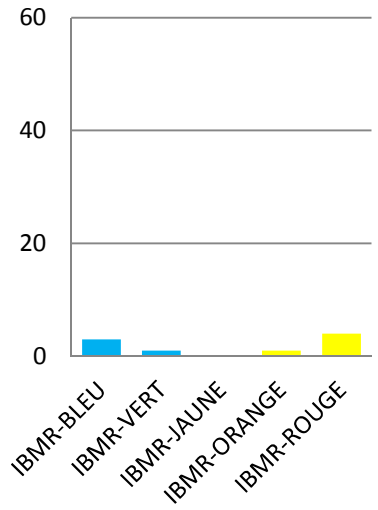
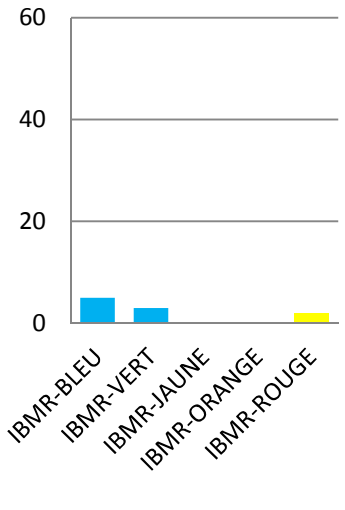
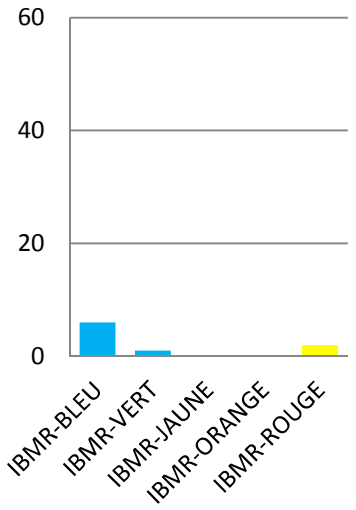
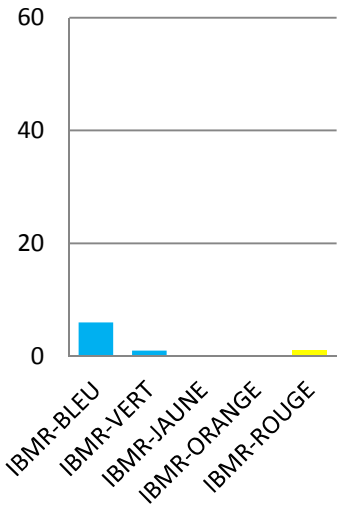
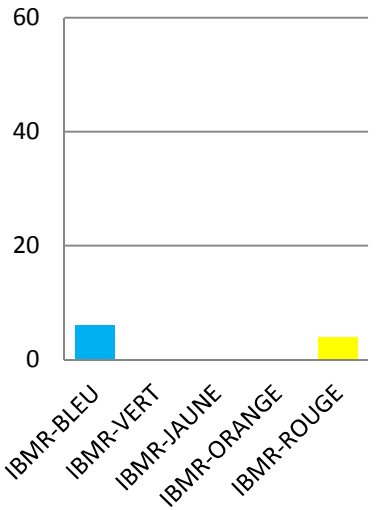
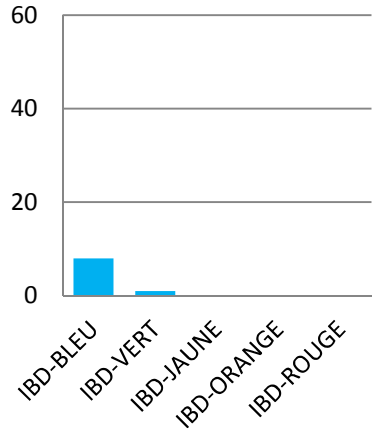
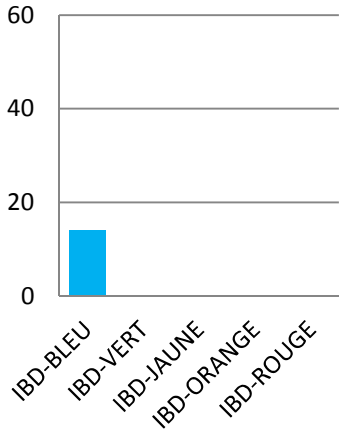
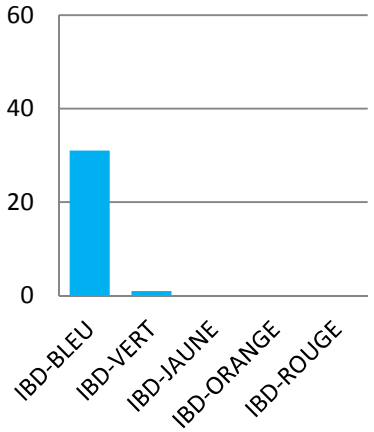
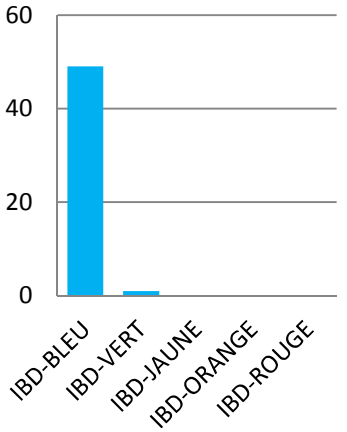
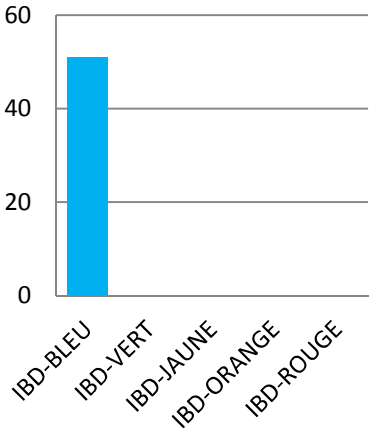
Altération NITR pour l'IBMR & l'IBD

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
IBMR					
IBD					

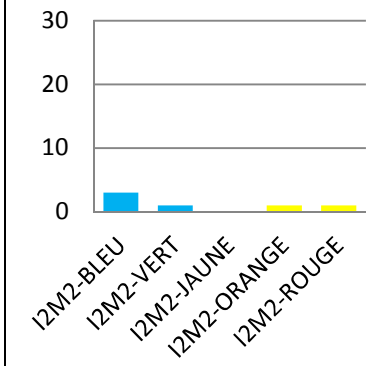
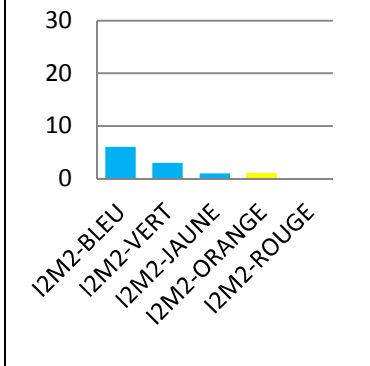
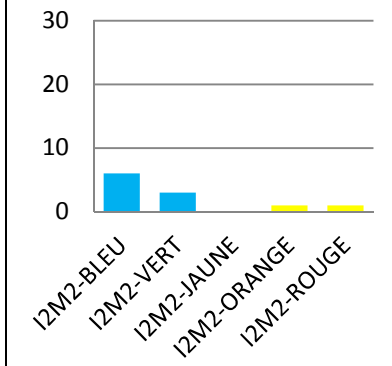
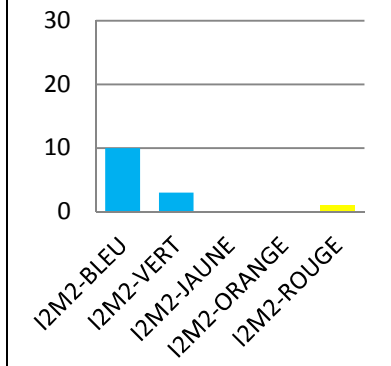
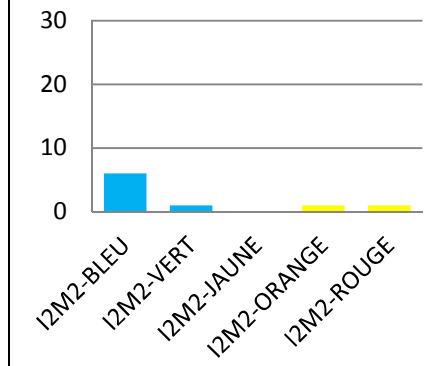
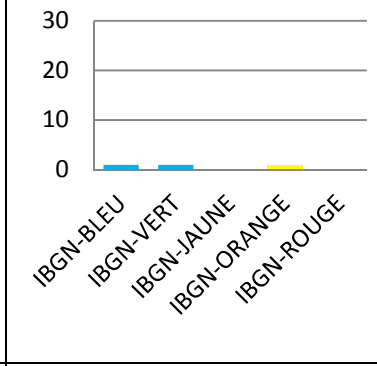
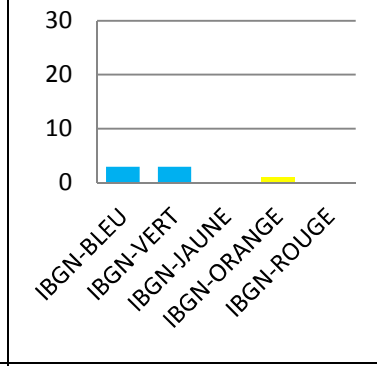
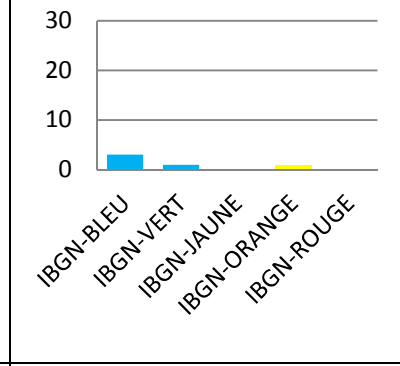
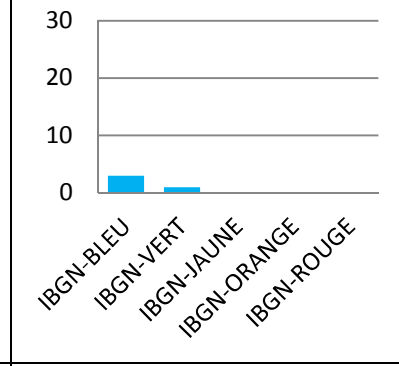
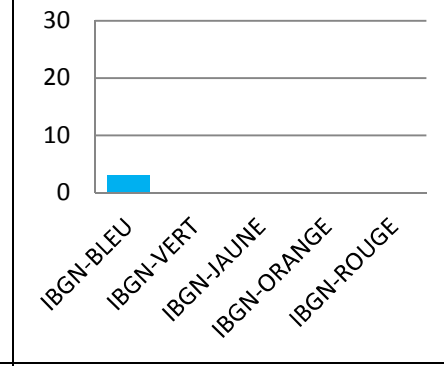
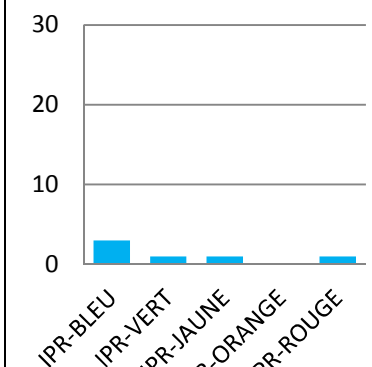
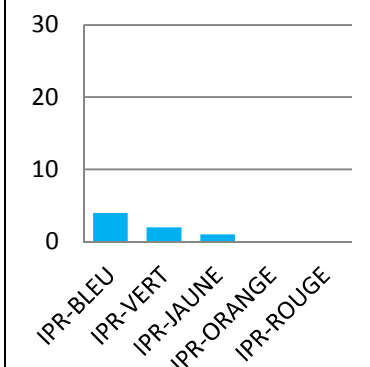
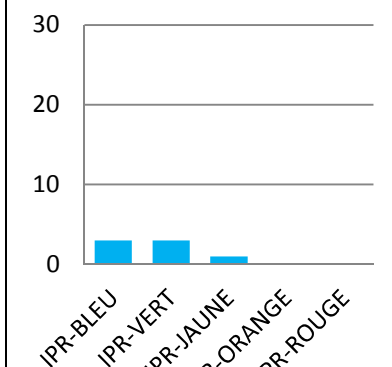
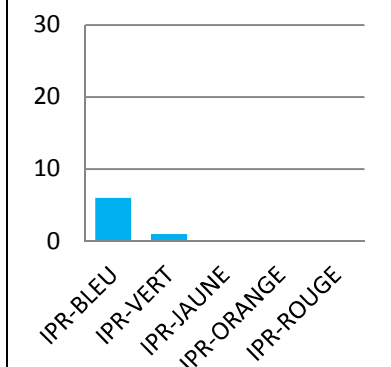
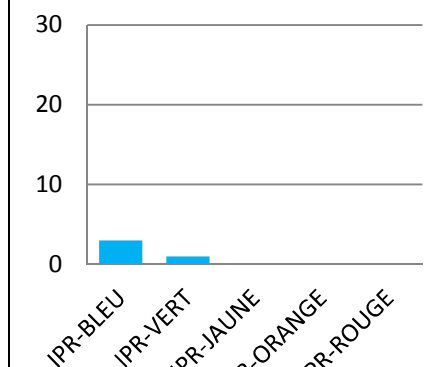
Altération **NITR** pour l'I2M2, l'IBGN & l'IPR

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
I2M2					
IBGN					
IPR					

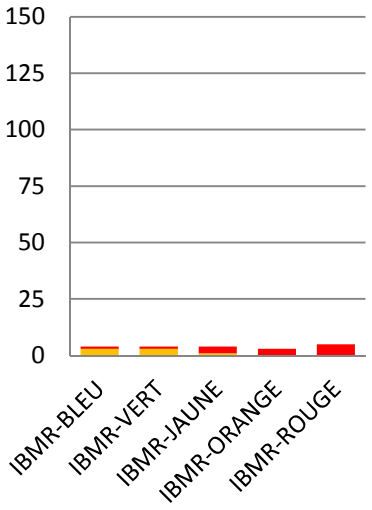
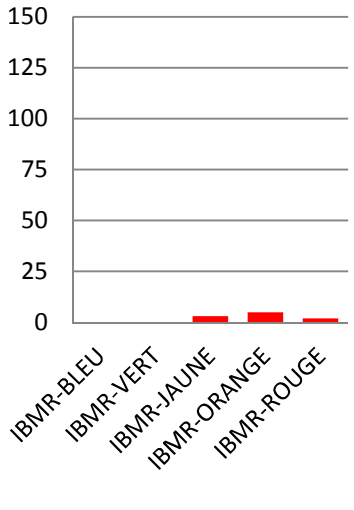
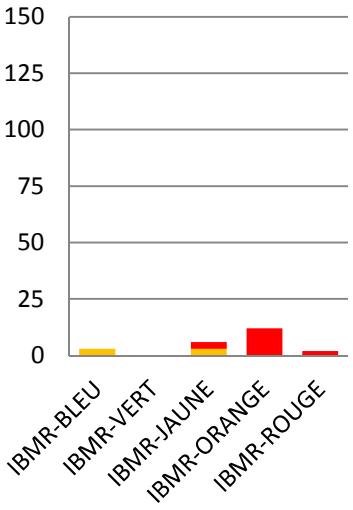
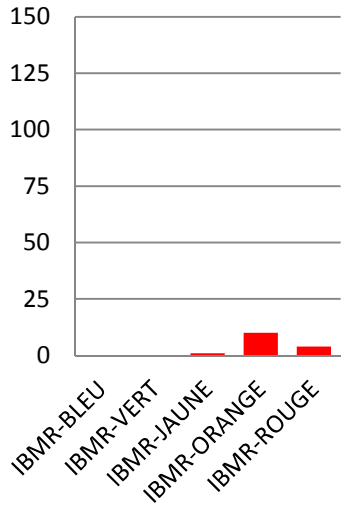
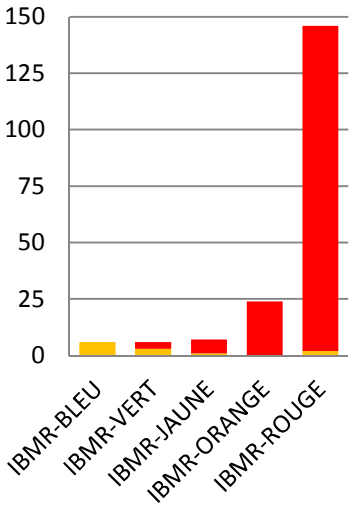
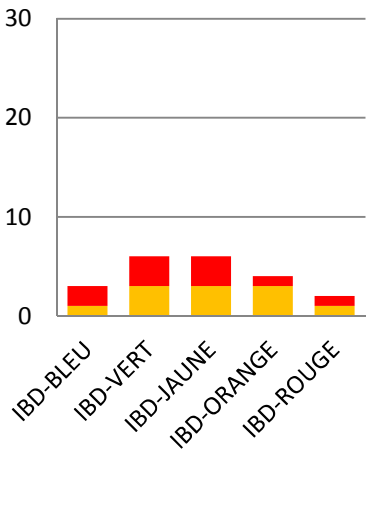
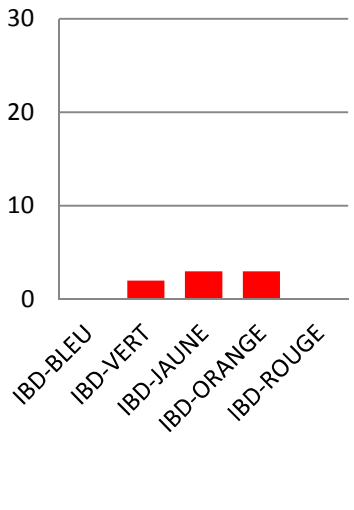
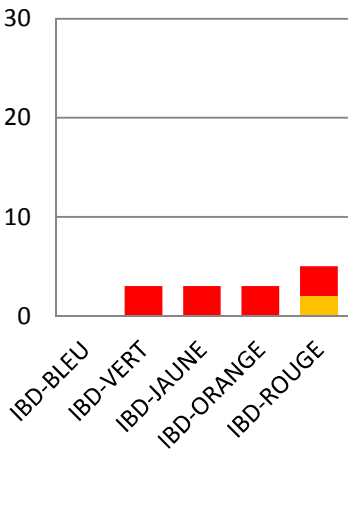
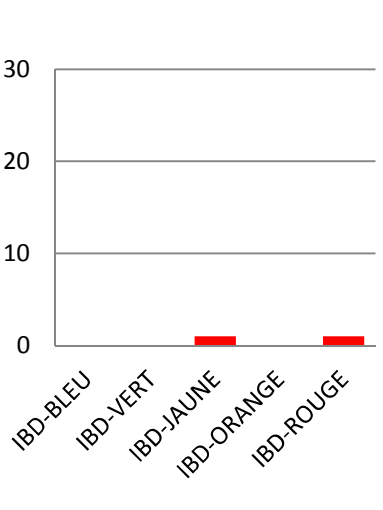
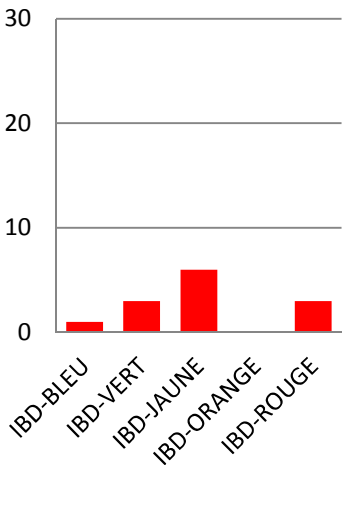
Altération PHOS pour l'IBMR & l'IBD

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
IBMR					
IBD					

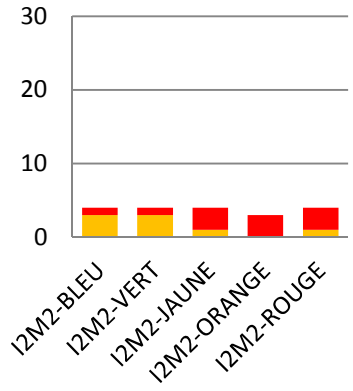
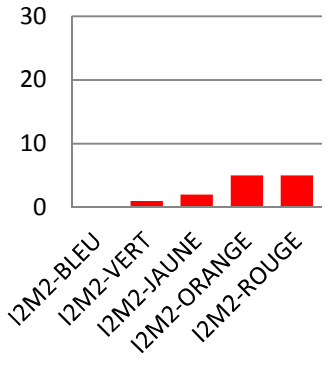
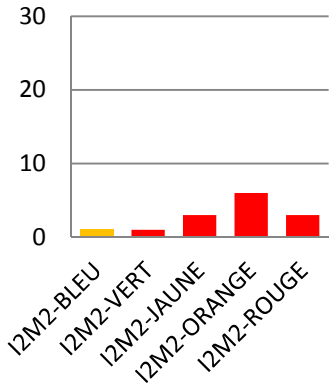
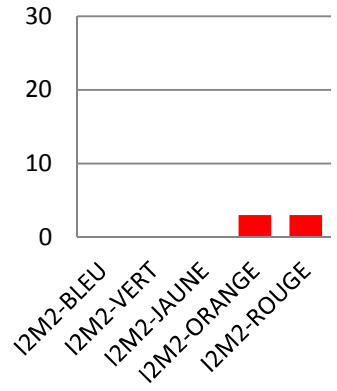
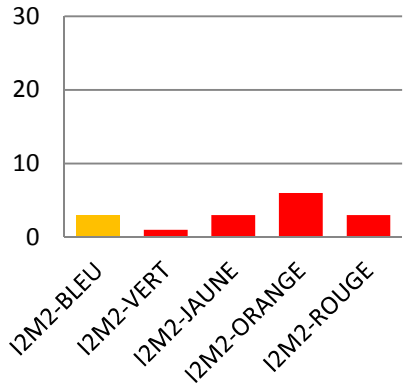
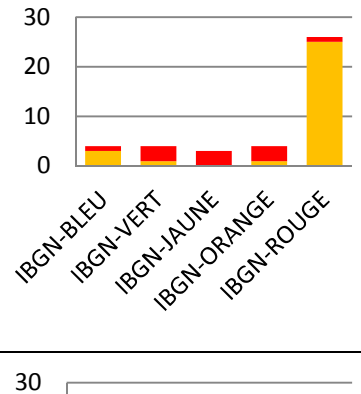
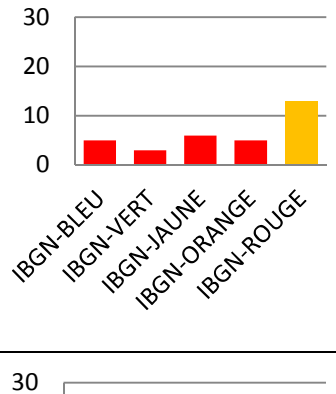
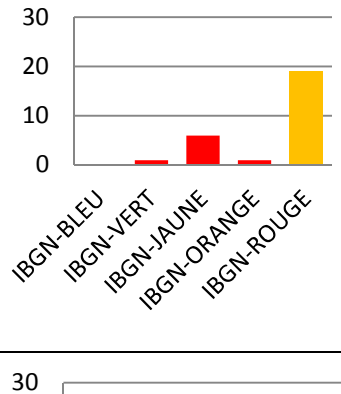
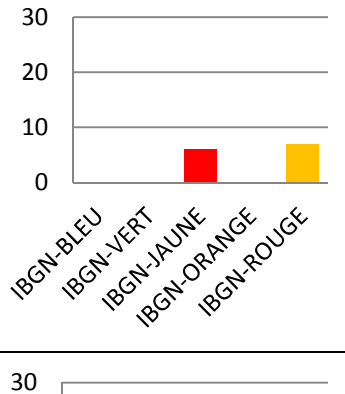
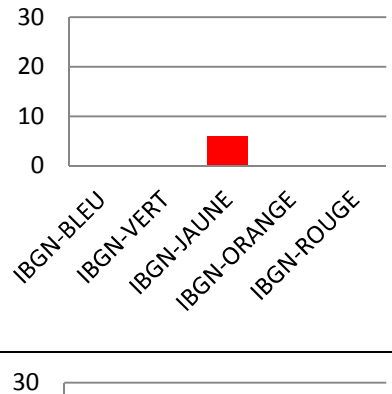
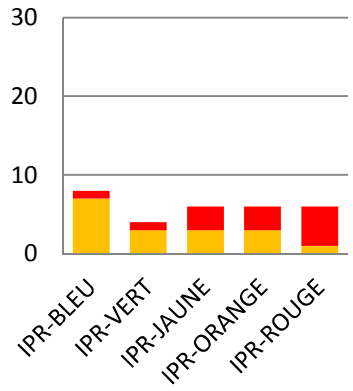
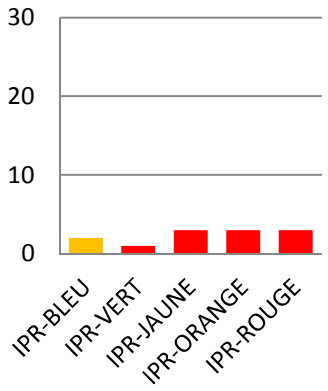
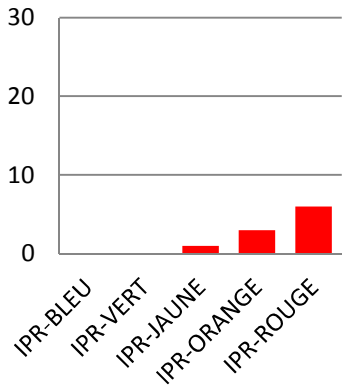
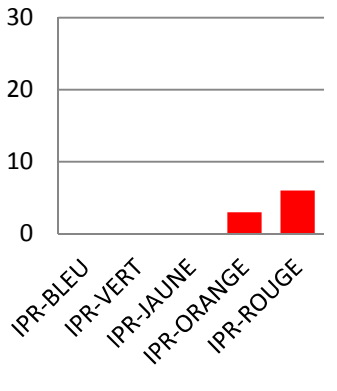
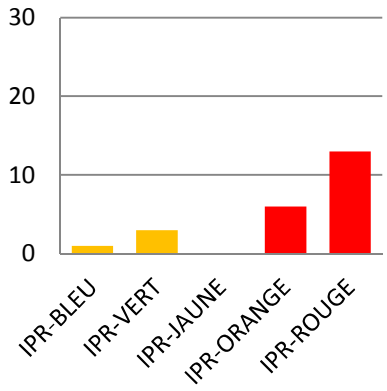
Altération **PHOS** pour l'I2M2, l'IBGN & l'IPR

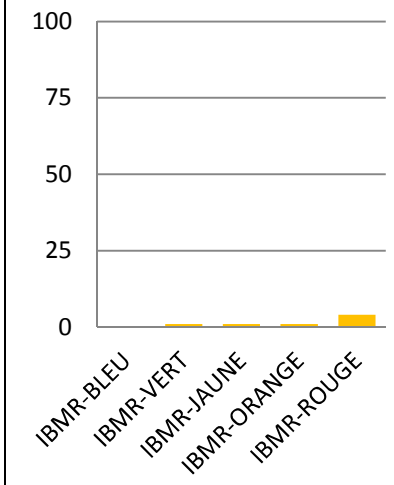
	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
I2M2					
IBGN					
IPR					

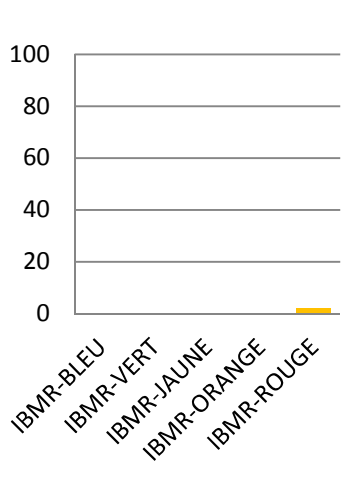
Altération PEST pour l'IBMR & l'IBD

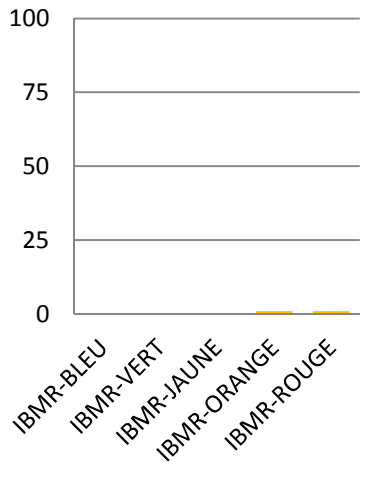
	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
IBMR					
IBD					

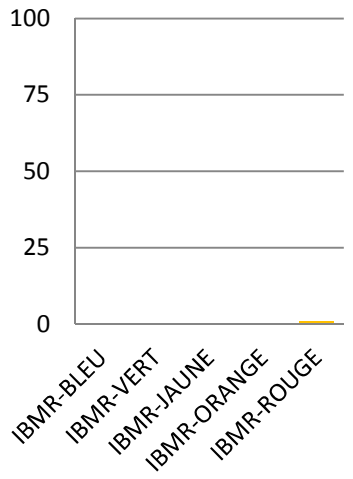
Altération PEST pour l'I2M2, l'IBGN & l'IPR

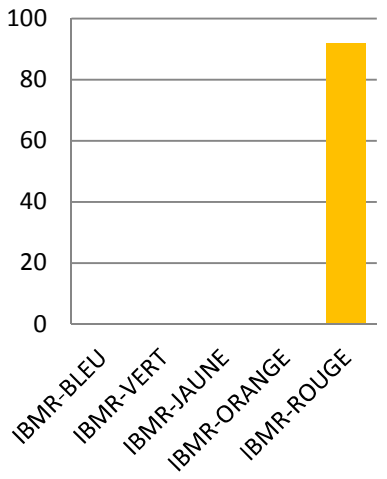
	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
I2M2					
IBGN					
IPR					

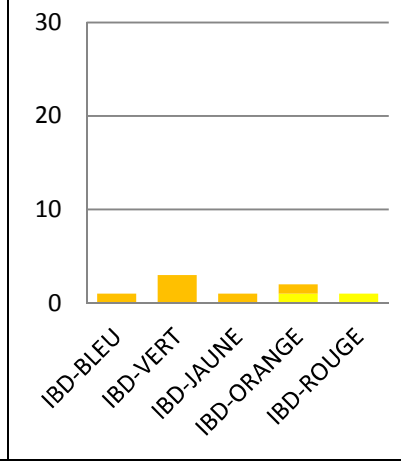


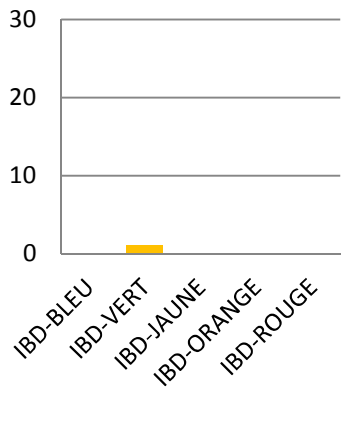


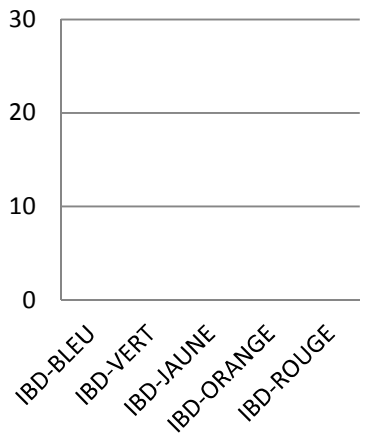


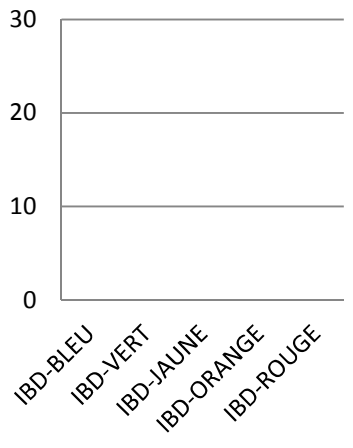


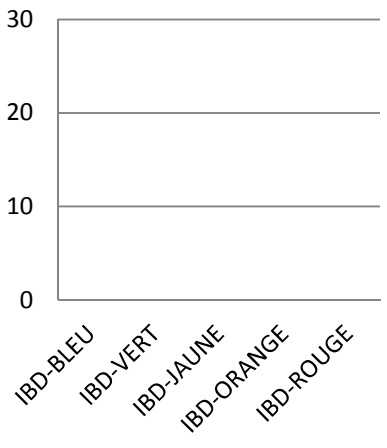








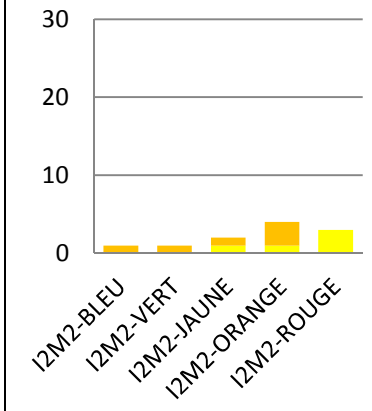
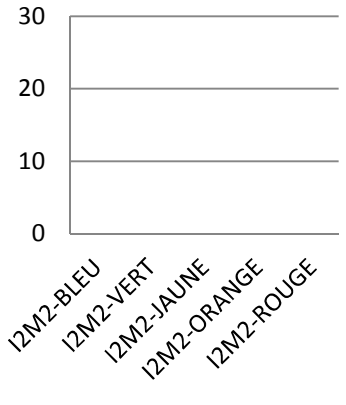
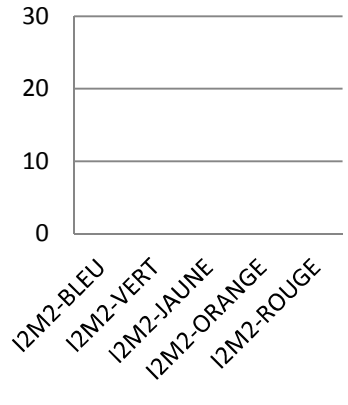
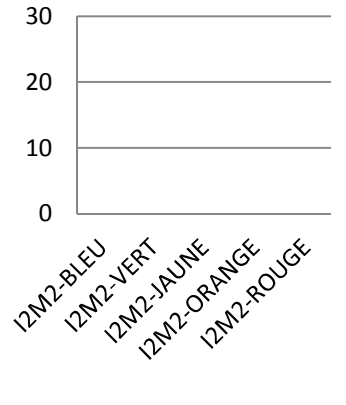
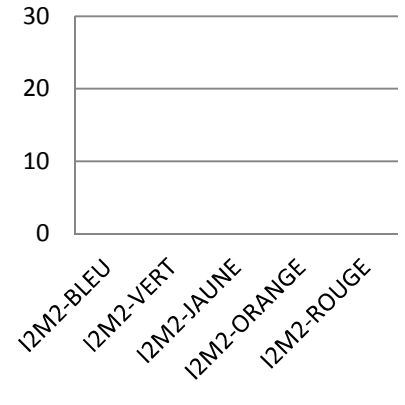
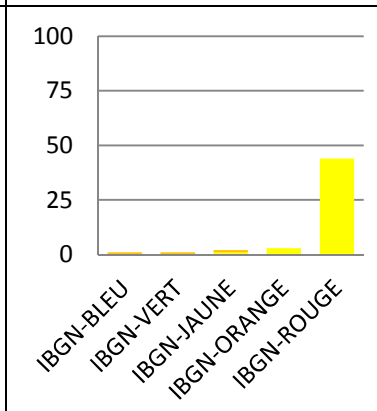
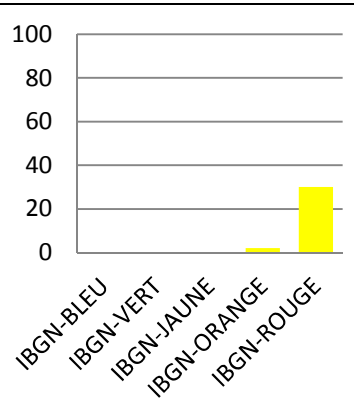
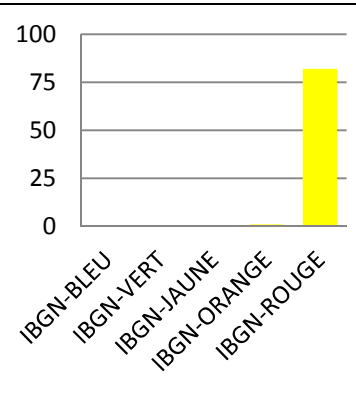
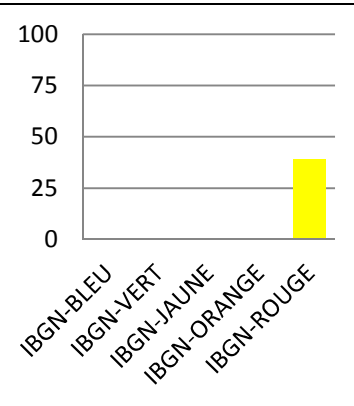
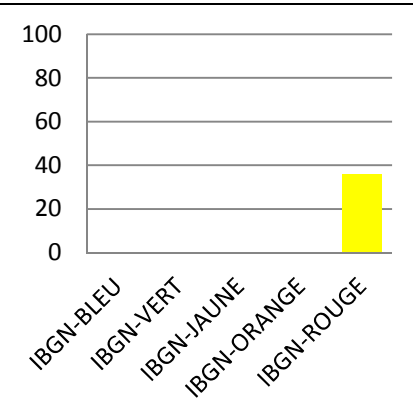
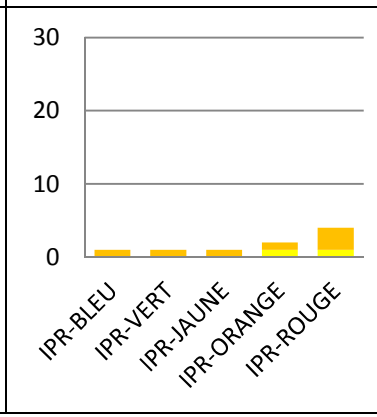
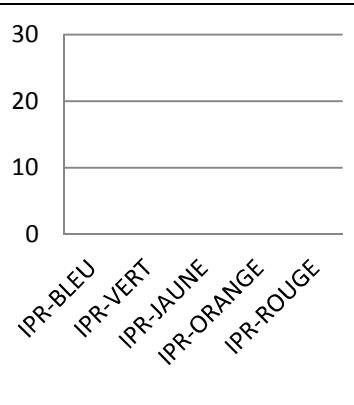
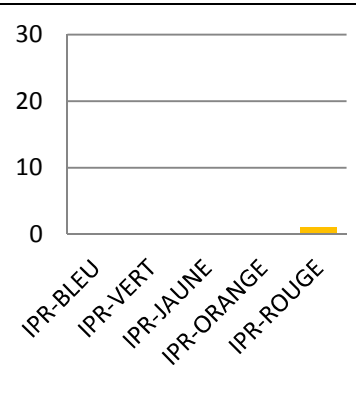
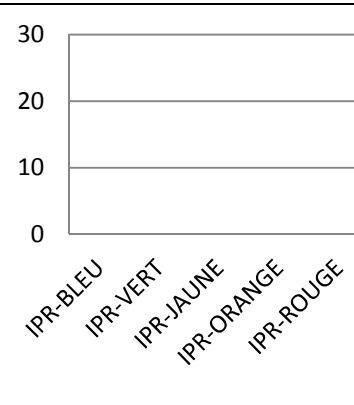
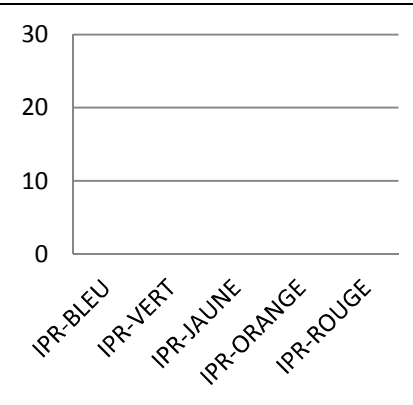




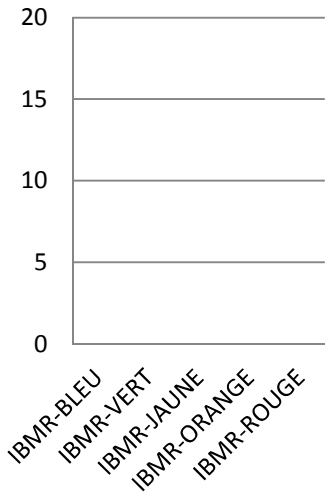
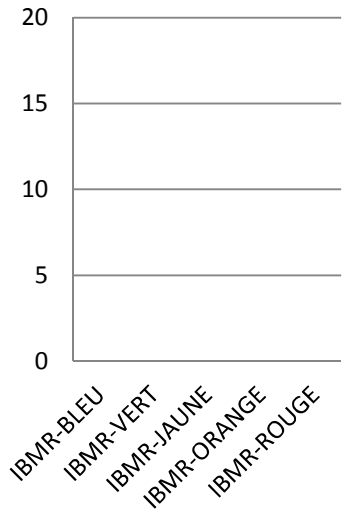
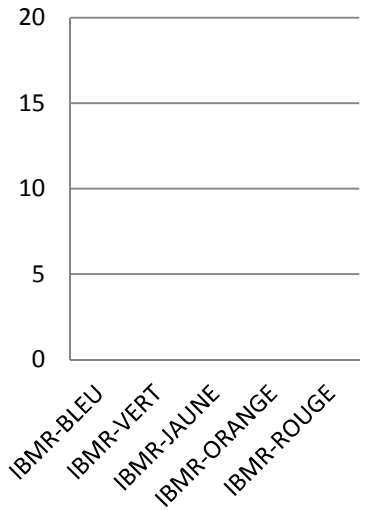
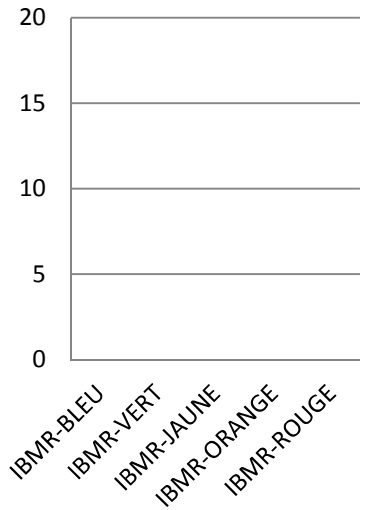
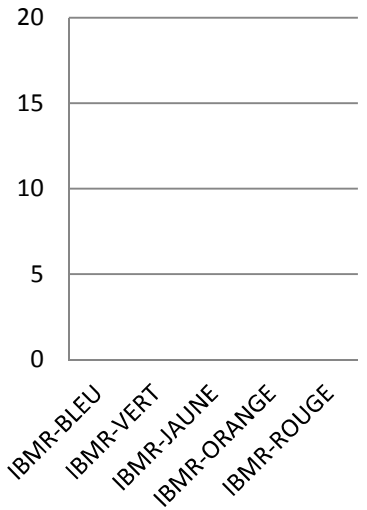
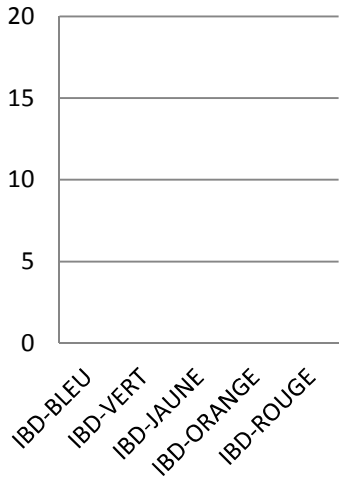
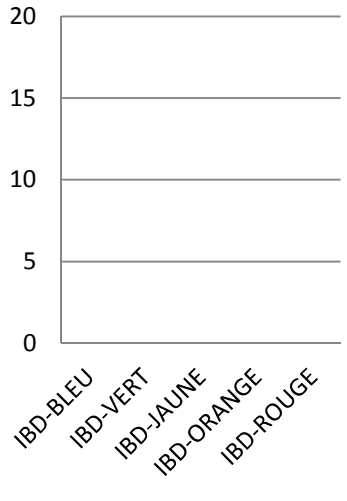
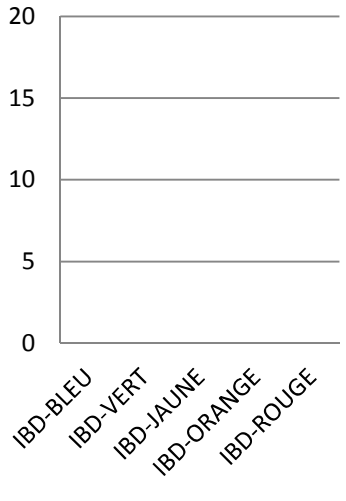
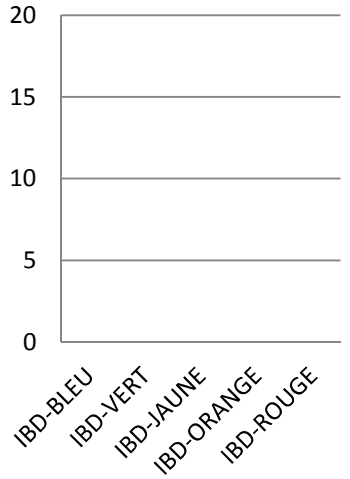
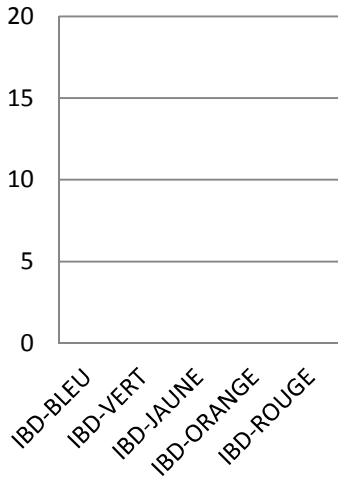
Altération **MPOR** pour l'IBMR & l'IBD

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
IBMR					
IBD					

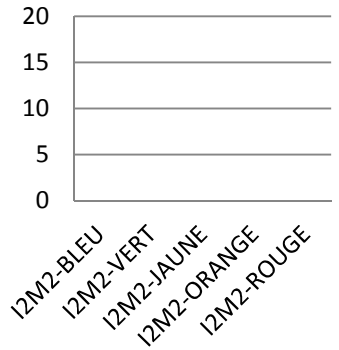
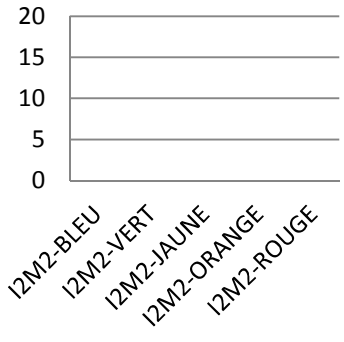
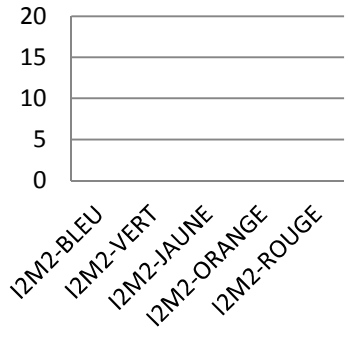
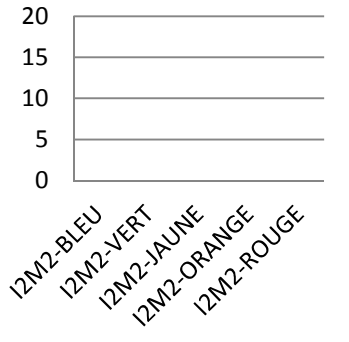
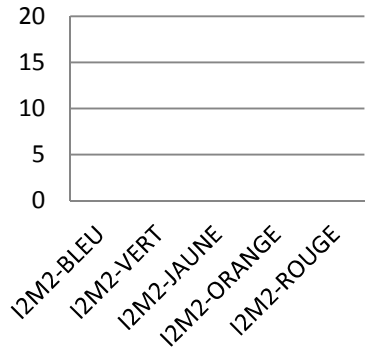
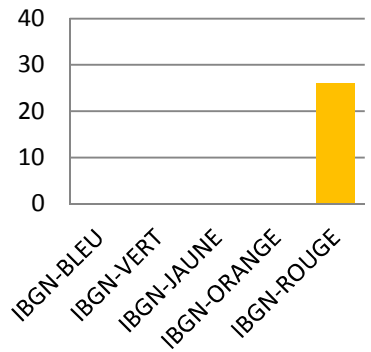
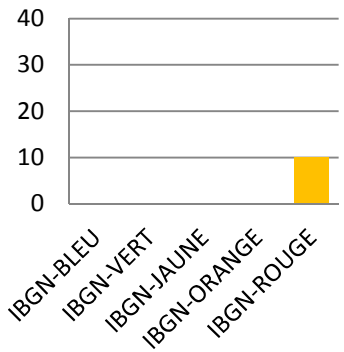
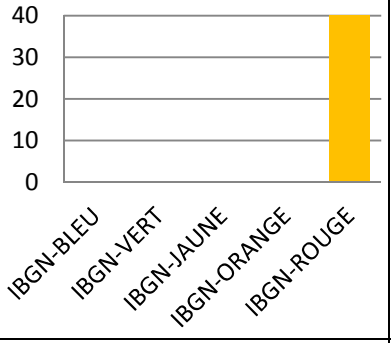
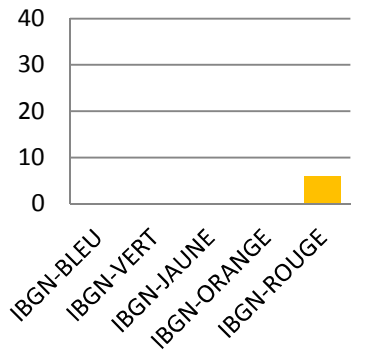
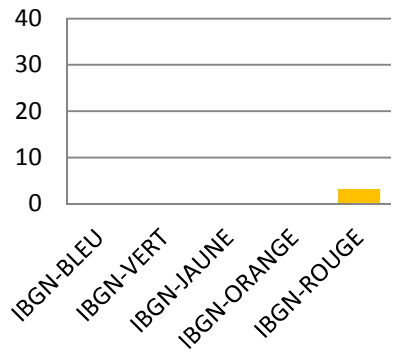
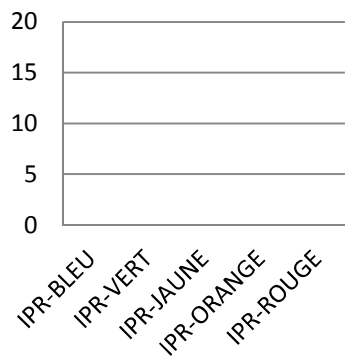
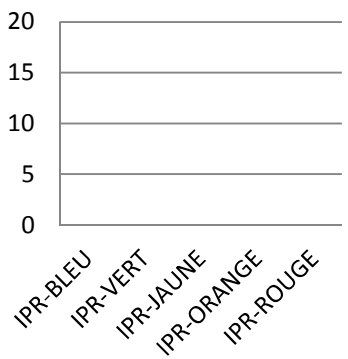
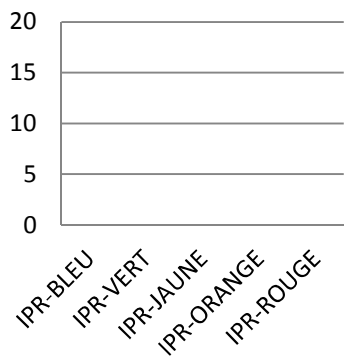
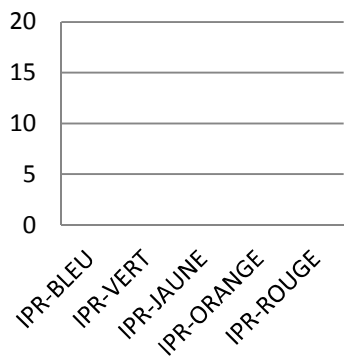
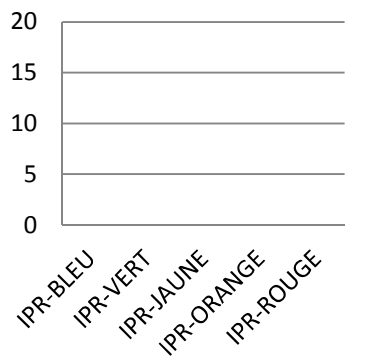
Altération MPOR pour l'I2M2, l'IBGN & l'IPR

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
I2M2					
IBGN					
IPR					

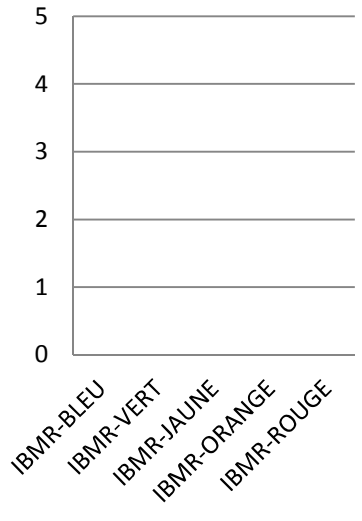
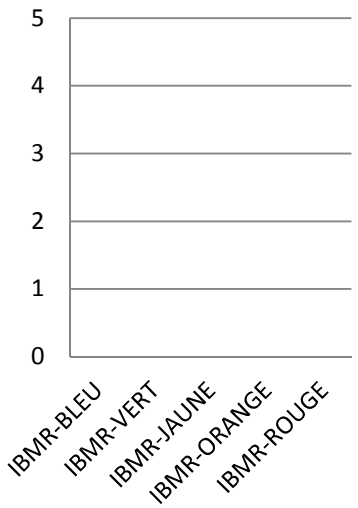
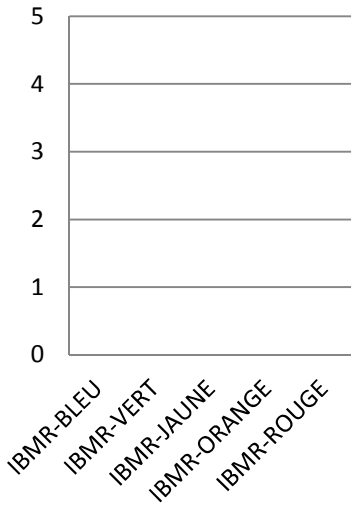
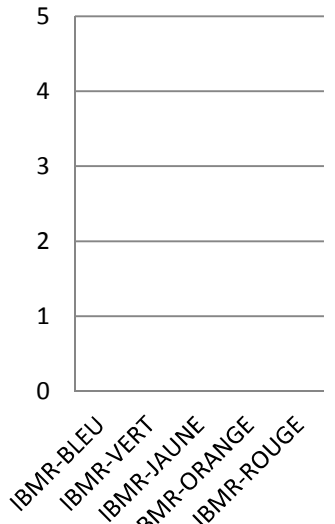
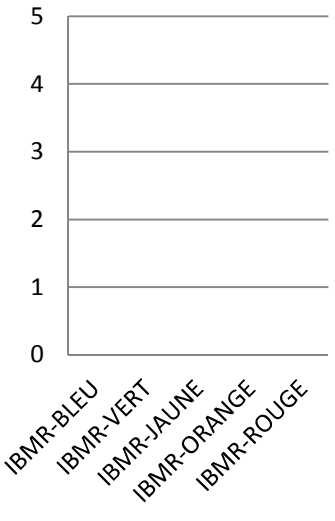
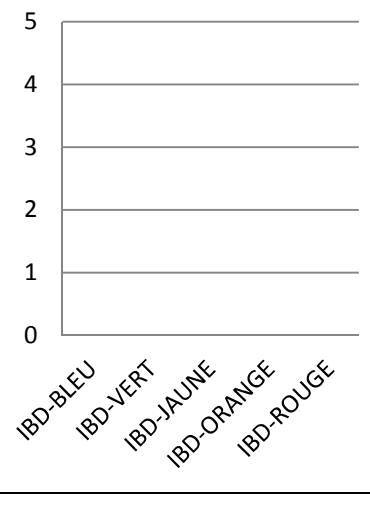
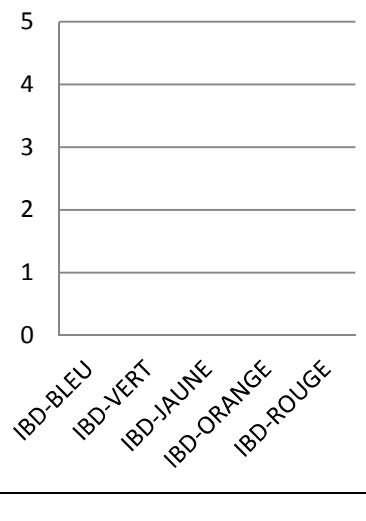
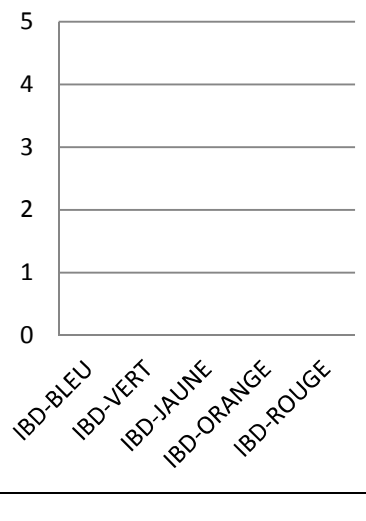
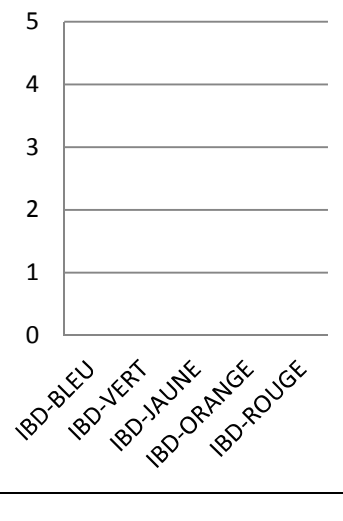
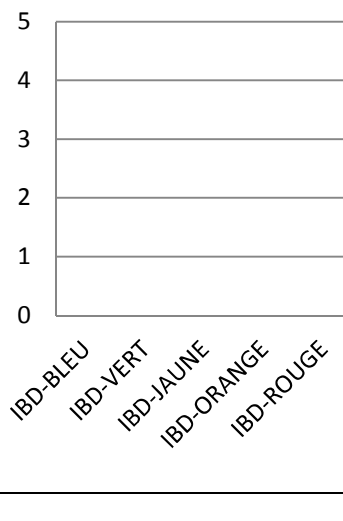
Altération MPMI pour l'IBMR & l'IBD

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
IBMR					
IBD					

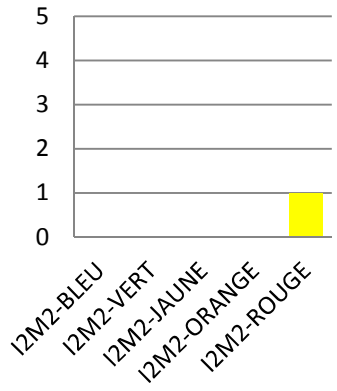
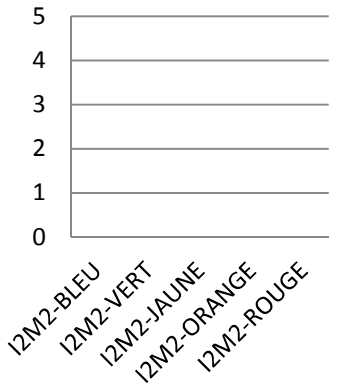
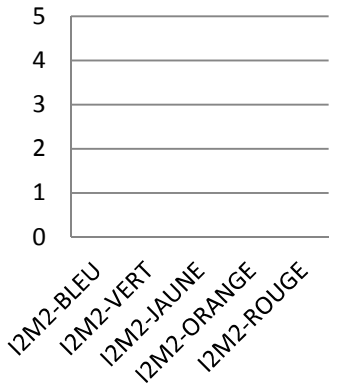
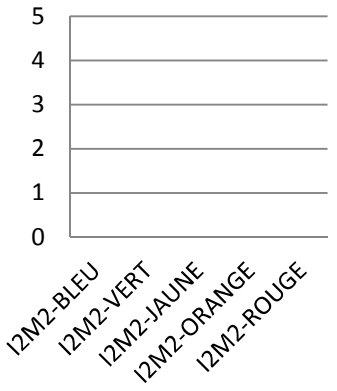
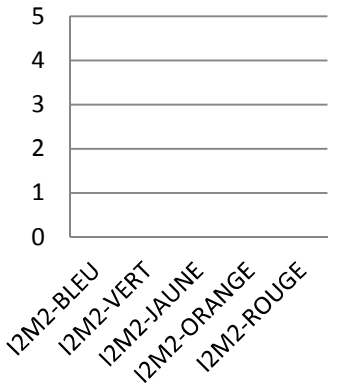
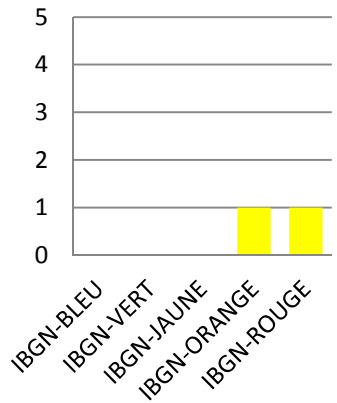
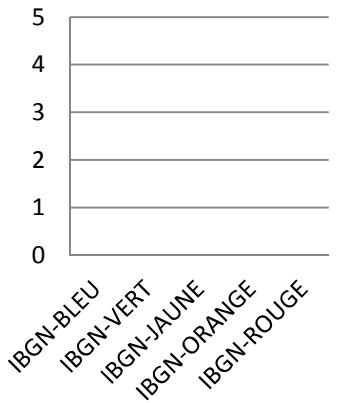
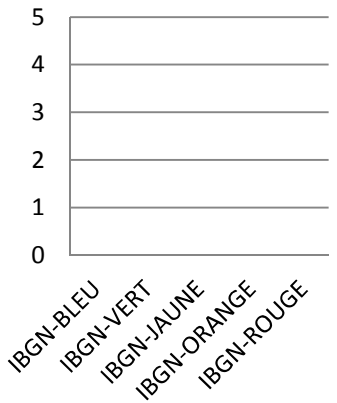
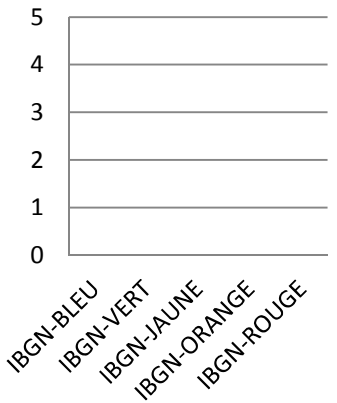
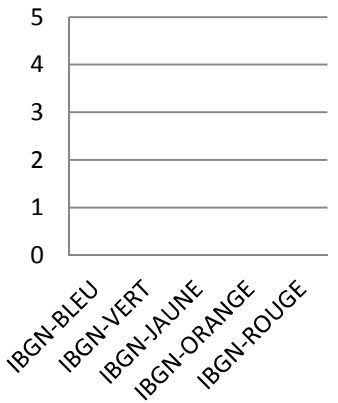
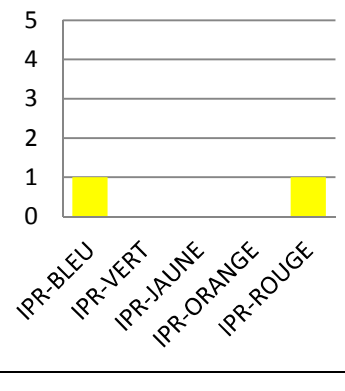
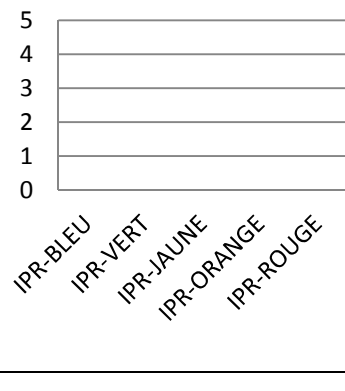
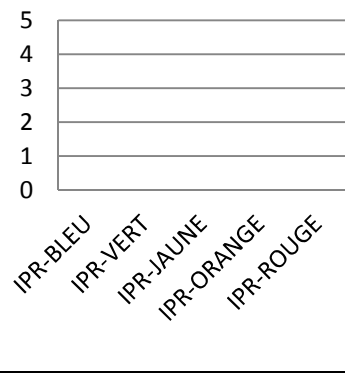
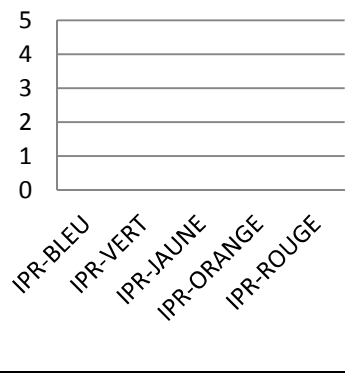
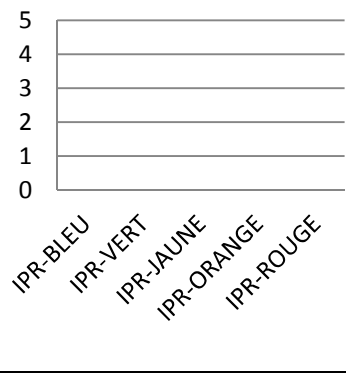
Altération **MPMI** pour l'I2M2, l'IBGN & l'IPR

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
I2M2					
IBGN					
IPR					

Altération PCB pour l’IBMR & l’IBD

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
IBMR					
IBD					

Altération PCB pour l'I2M2, l'IBGN & l'IPR

	3MOIS	6MOIS	12 MOIS	18 MOIS	24 MOIS
I2M2					
IBGN					
IPR					

ANNEXE 4 a: Caractéristiques des cinq motifs sélectionnés par contexte, sur la base de leur combinaison P la plus élevée, extraits pour la longueur de séquences 3 mois, avec le SEQ-eau

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E	Items du motif
60	IBD=Bleu	IBD=Bleu			0,349	0,333	0,960	1,689	1,801	1ACID_vert;1PHOS_Bleu;
58	IBD=Bleu	IBD=Bleu	2	1	0,389	0,333	0,960	1,642	1,766	1ACID_vert;1PHOS_Bleu;
61	IBD=Bleu	IBD=Bleu	2	2	0,386	0,333	0,960	1,631	1,754	1NITR_Bleu;1PHOS_Bleu;
59	IBD=Bleu	IBD=Bleu	2	1	0,396	0,333	0,960	1,607	1,734	1NITR_Bleu;1PHOS_Bleu;
57	IBD=Bleu	IBD=Bleu	2	2	0,465	0,333	0,840	1,505	1,635	2PHOS_Bleu;
93	IBD=Vert	IBD=Vert	2	2	0,317	0,333	0,880	1,055	1,148	2MPOR_Orange;
94	IBD=Vert	IBGN=Rouge	2	2	0,324	0,333	0,480		0,052	2PEST_Orange;
95	IBD=Vert	IBMR=Rouge	2	2	0,338	0,333	0,440		0,050	2PEST_Rouge;
86	IBD=Vert	IPR=Bleu	2	2	0,436	0,167	0,520		0,038	1MOOX_Bleu;
87	IBD=Vert	IBD=Bleu	1	1	0,377	0,167	0,520		0,033	1PHOS_Bleu;
71	IBD=Jaune	IBMR=Rouge	1	1	0,351	0,333	0,440		0,051	2PEST_Rouge;
70	IBD=Jaune	IBGN=Rouge	2	2	0,303	0,333	0,480		0,048	2PEST_Orange;
66	IBD=Jaune	IBMR=Rouge	2	2	0,345	0,167	0,560		0,032	1MOOX_Jaune;
64	IBD=Jaune	IBMR=Orange	1	1	0,467	0,167	0,160		0,012	1NITR_Jaune;
68	IBD=Jaune	IBMR=Rouge	1	1	0,367	0,167	0,200		0,012	1MPOR_Orange;
80	IBD=Orange	IBGN=Rouge	1	1	0,314	0,333	0,480		0,050	2PEST_Orange;
76	IBD=Orange	IBMR=Rouge	2	2	0,332	0,167	0,560		0,031	1MOOX_Jaune;
75	IBD=Orange	IBGN=Rouge	1	1	0,307	0,167	0,600		0,031	1MPOR_Jaune;
73	IBD=Orange	IBMR=Orange	1	1	0,472	0,167	0,160		0,013	1NITR_Jaune;
74	IBD=Orange	IBMR=Rouge	1	1	0,322	0,167	0,200		0,011	1MPOR_Orange;
83	IBD=Rouge	IBGN=Rouge	1	1	0,321	0,167	0,600		0,032	1MPOR_Jaune;
82	IBD=Rouge	IBGN=Rouge	1	1	0,321	0,167	0,160		0,009	1PEST_Orange;
84	IBD=Rouge	IBGN=Jaune	1	1	0,397	0,167	0,040		0,003	1PEST_Rouge;
8	I2M2=Bleu	IBD=Bleu	1	1	0,309	0,333	0,840		0,087	2PHOS_Bleu;
7	I2M2=Bleu	IBGN=Rouge	2	2	0,368	0,333	0,480		0,059	2PEST_Orange;
2	I2M2=Bleu	IBD=Bleu	2	2	0,584	0,167	0,520		0,051	1PHOS_Bleu;
1	I2M2=Bleu	IPR=Bleu	1	1	0,430	0,167	0,520		0,037	1MOOX_Bleu;
3	I2M2=Bleu	IBD=Bleu	1	1	0,355	0,167	0,200		0,012	1ACID_vert;
48	I2M2=Vert	IBGN=Rouge	1	1	0,303	0,333	0,480		0,048	2PEST_Orange;
47	I2M2=Vert	IBD=Bleu	2	2	0,468	0,167	0,520		0,041	1PHOS_Bleu;
42	I2M2=Vert	IPR=Bleu	1	1	0,351	0,167	0,520		0,030	1MOOX_Bleu;
43	I2M2=Vert	IBD=Bleu	1	1	0,404	0,167	0,200		0,013	1ACID_vert;
44	I2M2=Vert	IBMR=Rouge	1	1	0,366	0,167	0,200		0,012	1MPOR_Orange;
16	I2M2=Jaune	IBMR=Rouge	1	1	0,327	0,333	0,440		0,048	2PEST_Rouge;
11	I2M2=Jaune	IBGN=Rouge	2	2	0,309	0,167	0,600		0,031	1MPOR_Jaune;
13	I2M2=Jaune	IBMR=Rouge	1	1	0,356	0,167	0,200		0,012	1MPOR_Orange;
10	I2M2=Jaune	IBMR=Orange	1	1	0,439	0,167	0,160		0,012	1NITR_Jaune;
14	I2M2=Jaune	IBD=Bleu	1	1	0,317	0,167	0,200		0,011	1ACID_vert;
25	I2M2=Orange	IBD=Vert	1	1	0,300	0,333	0,880		0,088	2MPOR_Orange;
26	I2M2=Orange	IBMR=Rouge	2	2	0,399	0,333	0,440		0,059	2PEST_Rouge;
18	I2M2=Orange	IBMR=Rouge	2	2	0,348	0,167	0,800		0,046	1PHOS_Jaune;
19	I2M2=Orange	IBMR=Rouge	1	1	0,356	0,167	0,560		0,033	1MOOX_Jaune;
20	I2M2=Orange	IBGN=Rouge	1	1	0,318	0,167	0,600		0,032	1MPOR_Jaune;
29	I2M2=Rouge	I2M2=Rouge	1	1	0,388	0,167	0,840	1,528	1,582	1PCB_Jaune;
36	I2M2=Rouge	I2M2=Rouge	1	1	0,336	0,167	0,960	1,207	1,260	1NITR_Orange;
38	I2M2=Rouge	IBGN=Rouge	1	1	0,303	0,333	0,880		0,089	2MPOR_Jaune;
30	I2M2=Rouge	IBMR=Rouge	2	2	0,411	0,167	0,800		0,055	1PHOS_Jaune;
39	I2M2=Rouge	IBMR=Rouge	1	1	0,362	0,333	0,440		0,053	2PEST_Rouge;
104	IBGN=Bleu	IBGN=Rouge	2	2	0,335	0,333	0,480		0,054	2PEST_Orange;
97	IBGN=Bleu	IBD=Bleu	2	2	0,502	0,167	0,520		0,044	1PHOS_Bleu;
98	IBGN=Bleu	IPR=Bleu	1	1	0,406	0,167	0,520		0,035	1MOOX_Bleu;
102	IBGN=Bleu	IBD=Bleu	1	1	0,356	0,167	0,200		0,012	1ACID_vert;
99	IBGN=Bleu	IBMR=Rouge	1	1	0,352	0,167	0,200		0,012	1MPOR_Orange;
167	IBGN=Vert	IBMR=Rouge	1	1	0,316	0,333	0,440		0,046	2PEST_Rouge;
163	IBGN=Vert	IBD=Bleu	2	2	0,412	0,167	0,520		0,036	1PHOS_Bleu;
160	IBGN=Vert	IPR=Bleu	1	1	0,319	0,167	0,520		0,028	1MOOX_Bleu;
162	IBGN=Vert	IBD=Bleu	1	1	0,393	0,167	0,200		0,013	1ACID_vert;
161	IBGN=Vert	IBMR=Rouge	1	1	0,356	0,167	0,200		0,012	1MPOR_Orange;
110	IBGN=Jaune	IBGN=Jaune	1	1	0,524	0,167	0,040	1,035	1,038	1PEST_Rouge;
112	IBGN=Jaune	IBMR=Rouge	1	1	0,378	0,333	0,440		0,055	2PEST_Rouge;
108	IBGN=Jaune	IBGN=Rouge	2	2	0,351	0,167	0,600		0,035	1MPOR_Jaune;

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E	Items du motif
107	IBGN=Jaune	IBMR=Rouge	1	1	0,324	0,167	0,560		0,030	1MOOX_Jaune;
109	IBGN=Jaune	IBMR=Orange	1	1	0,422	0,167	0,160		0,011	1NITR_Jaune;
121	IBGN=Orange	IBGN=Rouge	1	1	0,305	0,333	0,880		0,090	2MPOR_Jaune;
122	IBGN=Orange	IBMR=Rouge	2	2	0,365	0,333	0,440		0,053	2PEST_Rouge;
115	IBGN=Orange	IBMR=Rouge	2	2	0,345	0,167	0,800		0,046	1PHOS_Jaune;
118	IBGN=Orange	IBGN=Rouge	1	1	0,433	0,167	0,600		0,043	1MPOR_Jaune;
116	IBGN=Orange	I2M2=Rouge	1	1	0,305	0,167	0,840		0,043	1PCB_Jaune;
150	IBGN=Rouge	IBGN=Rouge	1	1	0,333	0,833	0,960	3,071	3,337	1MPMI_Orange;2PEST_Orange;2MPOR_Jaune;
155	IBGN=Rouge	IBGN=Rouge	5	4	0,333	1,000	0,960	2,707	3,027	4PEST_Orange;2MPOR_Jaune;
145	IBGN=Rouge	IBGN=Rouge	6	6	0,444	0,667	0,960	2,734	3,018	2PEST_Orange;2MPOR_Jaune;
149	IBGN=Rouge	IBGN=Rouge	4	4	0,444	0,500	0,960	2,757	2,971	1MPMI_Orange;2MPOR_Jaune;
144	IBGN=Rouge	IBGN=Rouge	3	3	0,407	0,667	0,960	2,635	2,896	1MPMI_Orange;3MPOR_Jaune;
174	IBMR=Bleu	IBD=Bleu	4	3	0,331	0,333	0,840		0,093	2PHOS_Bleu;
173	IBMR=Bleu	IBGN=Rouge	2	2	0,370	0,333	0,480		0,059	2PEST_Orange;
172	IBMR=Bleu	IBD=Bleu	2	2	0,538	0,167	0,520		0,047	1PHOS_Bleu;
170	IBMR=Bleu	IPR=Bleu	1	1	0,468	0,167	0,520		0,041	1MOOX_Bleu;
169	IBMR=Bleu	IBGN=Rouge	1	1	0,399	0,167	0,160		0,011	1PEST_Orange;
214	IBMR=Vert	IBGN=Rouge	1	1	0,339	0,333	0,480		0,054	2PEST_Orange;
209	IBMR=Vert	IPR=Bleu	2	2	0,401	0,167	0,520		0,035	1MOOX_Bleu;
211	IBMR=Vert	IBD=Bleu	1	1	0,363	0,167	0,520		0,031	1PHOS_Bleu;
208	IBMR=Vert	IBMR=Rouge	1	1	0,321	0,167	0,200		0,011	1MPOR_Orange;
212	IBMR=Vert	IBGN=Rouge	1	1	0,366	0,167	0,160		0,010	1PEST_Orange;
183	IBMR=Jaune	IBMR=Rouge	1	1	0,332	0,333	0,440		0,049	2PEST_Rouge;
182	IBMR=Jaune	IBMR=Rouge	2	2	0,359	0,167	0,560		0,034	1MOOX_Jaune;
177	IBMR=Jaune	IPR=Bleu	1	1	0,328	0,167	0,520		0,028	1MOOX_Bleu;
176	IBMR=Jaune	IBMR=Orange	1	1	0,429	0,167	0,160		0,011	1NITR_Jaune;
179	IBMR=Jaune	IBMR=Rouge	1	1	0,338	0,167	0,200		0,011	1MPOR_Orange;
191	IBMR=Orange	IBMR=Orange	1	1	0,362	0,333	0,960	1,210	1,326	2NITR_Jaune;
186	IBMR=Orange	IBMR=Orange	2	2	0,601	0,167	0,160	1,169	1,185	1NITR_Jaune;
192	IBMR=Orange	IBMR=Rouge	1	1	0,383	0,333	0,440		0,056	2PEST_Rouge;
185	IBMR=Orange	IBMR=Rouge	2	2	0,319	0,167	0,800		0,043	1PHOS_Jaune;
187	IBMR=Orange	IBMR=Rouge	1	1	0,420	0,167	0,560		0,039	1MOOX_Jaune;
202	IBMR=Rouge	IBMR=Rouge	1	1	0,316	0,333	0,960	1,714	1,815	1MOOX_Jaune;1AZOT_Jaune;
204	IBMR=Rouge	IBMR=Rouge	2	2	0,316	0,333	0,960	1,684	1,785	1MOOX_Jaune;1PHOS_Jaune;
201	IBMR=Rouge	IBMR=Rouge	2	1	0,316	0,333	0,960	1,574	1,675	1AZOT_Jaune;1PHOS_Jaune;
203	IBMR=Rouge	IBMR=Rouge	2	2	0,368	0,333	0,960	1,534	1,652	1MOOX_Jaune;1PHOS_Jaune;
206	IBMR=Rouge	IBMR=Rouge	2	2	0,316	0,667	0,960	1,354	1,556	3PEST_Rouge;1MPOR_Orange;
227	IPR=Bleu	IPR=Bleu	4	3	0,401	0,333	0,960	1,448	1,577	2MOOX_Bleu;
216	IPR=Bleu	IPR=Bleu	2	2	0,346	0,167	0,960	1,361	1,416	1AZOT_Bleu;
225	IPR=Bleu	IPR=Bleu	1	1	0,303	0,333	0,960	1,249	1,346	1MOOX_Bleu;1PEST_Orange;
223	IPR=Bleu	IPR=Bleu	2	2	0,576	0,167	0,520	1,229	1,279	1MOOX_Bleu;
229	IPR=Bleu	IBGN=Rouge	1	1	0,303	0,500	0,920		0,139	3PEST_Orange;
271	IPR=Vert	IBGN=Rouge	3	3	0,344	0,333	0,480		0,055	2PEST_Orange;
267	IPR=Vert	IBD=Bleu	2	2	0,481	0,167	0,520		0,042	1PHOS_Bleu;
266	IPR=Vert	IPR=Bleu	1	1	0,422	0,167	0,520		0,037	1MOOX_Bleu;
264	IPR=Vert	IBD=Bleu	1	1	0,419	0,167	0,200		0,014	1ACID_vert;
265	IPR=Vert	IBMR=Rouge	1	1	0,343	0,167	0,200		0,011	1MPOR_Orange;
239	IPR=Jaune	IBGN=Rouge	1	1	0,307	0,333	0,480		0,049	2PEST_Orange;
238	IPR=Jaune	IBMR=Rouge	2	2	0,301	0,333	0,440		0,044	2PEST_Rouge;
234	IPR=Jaune	IBD=Bleu	2	2	0,361	0,167	0,520		0,031	1PHOS_Bleu;
237	IPR=Jaune	IPR=Bleu	1	1	0,337	0,167	0,520		0,029	1MOOX_Bleu;
233	IPR=Jaune	IBD=Bleu	1	1	0,361	0,167	0,200		0,012	1ACID_vert;
248	IPR=Orange	IBGN=Rouge	1	1	0,332	0,333	0,480		0,053	2PEST_Orange;
249	IPR=Orange	IBMR=Rouge	2	2	0,340	0,333	0,440		0,050	2PEST_Rouge;
242	IPR=Orange	IBGN=Rouge	2	2	0,310	0,167	0,600		0,031	1MPOR_Jaune;
241	IPR=Orange	IBMR=Rouge	1	1	0,318	0,167	0,560		0,030	1MOOX_Jaune;
245	IPR=Orange	IBMR=Rouge	1	1	0,374	0,167	0,200		0,012	1MPOR_Orange;
261	IPR=Rouge	IBMR=Rouge	1	1	0,315	0,333	0,960		0,101	1PEST_Rouge;1MPOR_Orange;
262	IPR=Rouge	IBMR=Rouge	2	2	0,313	0,333	0,960		0,100	1PEST_Rouge;1MPOR_Orange;
260	IPR=Rouge	IBMR=Rouge	2	1	0,354	0,333	0,440		0,052	2PEST_Rouge;
251	IPR=Rouge	I2M2=Rouge	2	2	0,321	0,167	0,840		0,045	1PCB_Jaune;
255	IPR=Rouge	IBGN=Rouge	1	1	0,307	0,167	0,600		0,031	1MPOR_Jaune;

ANNEXE 4 b: Caractéristiques des cinq motifs sélectionnés par contexte, sur la base de leur combinaison P la plus élevée, extraits pour la longueur de séquences 6 mois, avec le SEQ-eau

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E	Items du motif
58	IBD=Bleu	IBD=Bleu	4	3	0,412	0,333	0,960	1,638	1,769	1NITR_Bleu;3PHOS_Bleu;
71	IBD=Bleu	IBD=Bleu	5	4	0,413	0,417	0,960	1,539	1,704	2ACID_vert;3PHOS_Bleu;
72	IBD=Bleu	IBD=Bleu	5	4	0,431	0,417	0,960	1,532	1,704	2ACID_vert;3PHOS_Bleu;
65	IBD=Bleu	IBD=Bleu	4	4	0,406	0,333	0,960	1,570	1,700	1ACID_vert;3PHOS_Bleu;
60	IBD=Bleu	IBD=Bleu	3	3	0,425	0,250	0,960	1,595	1,697	1NITR_Bleu;2PHOS_Bleu;
102	IBD=Vert	IPR=Bleu	2	2	0,437	0,167	0,800		0,058	2MOOX_Bleu;
101	IBD=Vert	IBMR=Orange	2	2	0,409	0,167	0,520		0,035	2NITR_Jaune;
95	IBD=Vert	IPR=Rouge	1	1	0,412	0,083	0,880		0,030	1MPOR_Orange;
100	IBD=Vert	IBGN=Jaune	2	2	0,419	0,167	0,360		0,025	2PEST_Rouge;
96	IBD=Vert	IBD=Bleu	1	1	0,516	0,083	0,520		0,022	1PHOS_Bleu;
79	IBD=Jaune	IBMR=Orange	2	2	0,487	0,167	0,520		0,042	2NITR_Jaune;
74	IBD=Jaune	IBD=Orange	1	1	0,407	0,083	0,840		0,028	1MINE_Vert;
80	IBD=Jaune	IBGN=Jaune	2	2	0,448	0,167	0,360		0,027	2PEST_Rouge;
75	IBD=Jaune	IBMR=Rouge	1	1	0,460	0,083	0,520		0,020	1MOOX_Jaune;
76	IBD=Jaune	IPR=Rouge	1	1	0,471	0,083	0,320		0,013	1PEST_Rouge;
86	IBD=Orange	IBD=Orange	1	1	0,500	0,083	0,840	1,000	1,035	1MINE_Vert;
88	IBD=Orange	IBMR=Orange	2	2	0,498	0,167	0,520		0,043	2NITR_Jaune;
87	IBD=Orange	IBGN=Jaune	2	2	0,428	0,167	0,360		0,026	2PEST_Rouge;
84	IBD=Orange	IBMR=Rouge	1	1	0,438	0,083	0,520		0,019	1MOOX_Jaune;
82	IBD=Orange	IPR=Rouge	1	1	0,454	0,083	0,320		0,012	1PEST_Rouge;
90	IBD=Rouge	IBD=Rouge	1	1	0,500	0,083	0,840	1,000	1,035	1MINE_Vert;
91	IBD=Rouge	IBMR=Orange	1	1	0,423	0,083	0,160		0,006	1NITR_Jaune;
92	IBD=Rouge	IBD=Bleu	1	1	0,474	0,083	0,000		0,000	1ACID_vert;
107	IBGN=Bleu	IBD=Bleu	2	2	0,518	0,167	0,640		0,055	2PHOS_Bleu;
106	IBGN=Bleu	IBD=Bleu	1	1	0,625	0,083	0,520		0,027	1PHOS_Bleu;
104	IBGN=Bleu	IPR=Bleu	1	1	0,480	0,083	0,520		0,021	1MOOX_Bleu;
105	IBGN=Bleu	IBD=Bleu	1	1	0,511	0,083	0,000		0,000	1ACID_vert;
185	IBGN=Vert	IBD=Bleu	2	2	0,416	0,167	0,640		0,044	2PHOS_Bleu;
184	IBGN=Vert	IBGN=Jaune	2	2	0,410	0,167	0,360		0,025	2PEST_Rouge;
180	IBGN=Vert	IBD=Bleu	1	1	0,537	0,083	0,520		0,023	1PHOS_Bleu;
183	IBGN=Vert	IPR=Bleu	1	1	0,407	0,083	0,520		0,018	1MOOX_Bleu;
179	IBGN=Vert	IPR=Rouge	1	1	0,437	0,083	0,320		0,012	1PEST_Rouge;
115	IBGN=Jaune	IBGN=Jaune	3	3	0,444	0,250	0,800	1,003	1,091	3PEST_Rouge;
114	IBGN=Jaune	IBGN=Jaune	2	2	0,506	0,167	0,360	1,032	1,063	2PEST_Rouge;
113	IBGN=Jaune	IBMR=Orange	2	2	0,438	0,167	0,520		0,038	2NITR_Jaune;
112	IBGN=Jaune	IBMR=Rouge	1	1	0,424	0,083	0,520		0,018	1MOOX_Jaune;
109	IBGN=Jaune	IPR=Rouge	1	1	0,538	0,083	0,320		0,014	1PEST_Rouge;
127	IBGN=Orange	IBGN=Jaune	3	3	0,415	0,250	0,800		0,083	3PEST_Rouge;
124	IBGN=Orange	IBGN=Rouge	2	2	0,406	0,167	0,960		0,065	2MPOR_Jaune;
120	IBGN=Orange	IBGN=Rouge	1	1	0,435	0,083	0,960		0,035	1MPOR_Jaune;
126	IBGN=Orange	IBMR=Orange	2	2	0,401	0,167	0,520		0,035	2NITR_Jaune;
117	IBGN=Orange	I2M2=Rouge	1	1	0,454	0,083	0,800		0,030	1PHOS_Jaune;
169	IBGN=Rouge	IBGN=Rouge	12	12	0,444	1,000	0,960	3,538	3,965	7PEST_Orange;5MPOR_Jaune;
170	IBGN=Rouge	IBGN=Rouge	10	10	0,407	0,833	0,960	3,486	3,812	1NITR_Jaune;6MPOR_Jaune;3MPMI_Orange;
167	IBGN=Rouge	IBGN=Rouge	12	6	0,407	1,000	0,960	3,373	3,764	6PEST_Orange;6MPOR_Jaune;
171	IBGN=Rouge	IBGN=Rouge	7	7	0,444	0,583	0,960	2,976	3,225	5MPOR_Jaune;2MPMI_Orange;
162	IBGN=Rouge	IBGN=Rouge	10	8	0,407	0,833	0,960	2,852	3,178	7MPOR_Jaune;3MPMI_Orange;
7	I2M2=Bleu	IBD=Bleu	3	3	0,461	0,250	0,840		0,097	3PHOS_Bleu;
6	I2M2=Bleu	IBD=Bleu	2	2	0,610	0,167	0,640		0,065	2PHOS_Bleu;
5	I2M2=Bleu	IPR=Bleu	2	2	0,416	0,167	0,800		0,055	2MOOX_Bleu;
1	I2M2=Bleu	IBMR=Bleu	1	1	0,433	0,083	0,920		0,033	1MINE_Orange;
2	I2M2=Bleu	IBD=Bleu	1	1	0,715	0,083	0,520		0,031	1PHOS_Bleu;
43	I2M2=Vert	IBD=Bleu	2	2	0,481	0,167	0,640		0,051	2PHOS_Bleu;
41	I2M2=Vert	IBD=Bleu	1	1	0,610	0,083	0,520		0,026	1PHOS_Bleu;
38	I2M2=Vert	IPR=Bleu	1	1	0,432	0,083	0,520		0,019	1MOOX_Bleu;
40	I2M2=Vert	IPR=Rouge	1	1	0,426	0,083	0,320		0,011	1PEST_Rouge;
42	I2M2=Vert	IBMR=Orange	1	1	0,401	0,083	0,160		0,005	1NITR_Jaune;
14	I2M2=Jaune	IBMR=Orange	2	2	0,423	0,167	0,520		0,037	2NITR_Jaune;
13	I2M2=Jaune	IBGN=Jaune	2	2	0,429	0,167	0,360		0,026	2PEST_Rouge;
10	I2M2=Jaune	IBD=Bleu	1	1	0,400	0,083	0,520		0,017	1PHOS_Bleu;
12	I2M2=Jaune	IPR=Rouge	1	1	0,451	0,083	0,320		0,012	1PEST_Rouge;
11	I2M2=Jaune	IBMR=Orange	1	1	0,560	0,083	0,160		0,007	1NITR_Jaune;
24	I2M2=Orange	IBGN=Jaune	3	3	0,443	0,250	0,800		0,089	3PEST_Rouge;

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E	Items du motif
22	I2M2=Orange	IBMR=Orange	2	2	0,527	0,167	0,520		0,046	2NITR_Jaune;
17	I2M2=Orange	I2M2=Rouge	1	1	0,415	0,083	0,880		0,030	1NITRE_Orange;
23	I2M2=Orange	IBGN=Jaune	2	2	0,490	0,167	0,360		0,029	2PEST_Rouge;
16	I2M2=Orange	I2M2=Rouge	1	1	0,436	0,083	0,800		0,029	1PHOS_Jaune;
26	I2M2=Rouge	I2M2=Rouge	1	1	0,529	0,083	0,960	1,467	1,509	1AZOT_Jaune;
30	I2M2=Rouge	I2M2=Rouge	1	1	0,552	0,083	0,800	1,266	1,303	1PHOS_Jaune;
28	I2M2=Rouge	I2M2=Rouge	1	1	0,455	0,083	0,880	1,094	1,128	1NITRE_Orange;
36	I2M2=Rouge	IBGN=Jaune	3	3	0,409	0,250	0,800		0,082	3PEST_Rouge;
34	I2M2=Rouge	IBMR=Orange	2	2	0,497	0,167	0,520		0,043	2NITR_Jaune;
187	IBMR=Bleu	IBMR=Bleu	1	1	0,450	0,083	0,920	1,041	1,075	1MINE_Orange;
193	IBMR=Bleu	IBD=Bleu	3	3	0,459	0,250	0,840		0,096	3PHOS_Bleu;
191	IBMR=Bleu	IPR=Bleu	2	2	0,459	0,167	0,800		0,061	2MOOX_Bleu;
192	IBMR=Bleu	IBD=Bleu	2	2	0,566	0,167	0,640		0,060	2PHOS_Bleu;
190	IBMR=Bleu	IBD=Bleu	1	1	0,664	0,083	0,520		0,029	1PHOS_Bleu;
235	IBMR=Vert	IBD=Bleu	2	2	0,402	0,167	0,640		0,043	2PHOS_Bleu;
233	IBMR=Vert	IPR=Bleu	1	1	0,512	0,083	0,520		0,022	1MOOX_Bleu;
232	IBMR=Vert	IBD=Bleu	1	1	0,510	0,083	0,520		0,022	1PHOS_Bleu;
234	IBMR=Vert	IBMR=Orange	1	1	0,422	0,083	0,160		0,006	1NITR_Jaune;
231	IBMR=Vert	IBD=Bleu	1	1	0,480	0,083	0,000		0,000	1ACID_vert;
201	IBMR=Jaune	IBMR=Orange	2	2	0,459	0,167	0,520		0,040	2NITR_Jaune;
200	IBMR=Jaune	IBGN=Jaune	2	2	0,417	0,167	0,360		0,025	2PEST_Rouge;
195	IBMR=Jaune	IBMR=Rouge	1	1	0,469	0,083	0,520		0,020	1MOOX_Jaune;
197	IBMR=Jaune	IPR=Bleu	1	1	0,425	0,083	0,520		0,018	1MOOX_Bleu;
196	IBMR=Jaune	IPR=Rouge	1	1	0,436	0,083	0,320		0,012	1PEST_Rouge;
208	IBMR=Orange	IBMR=Orange	2	2	0,439	0,167	0,960	1,169	1,240	1ACID_vert;1NITR_Jaune;
213	IBMR=Orange	IBMR=Orange	3	3	0,423	0,250	0,960	1,136	1,238	3NITR_Jaune;
211	IBMR=Orange	IBMR=Orange	2	2	0,577	0,167	0,520	1,094	1,144	2NITR_Jaune;
207	IBMR=Orange	IBMR=Orange	1	1	0,735	0,083	0,160	1,095	1,105	1NITR_Jaune;
212	IBMR=Orange	IBGN=Jaune	3	3	0,413	0,250	0,800		0,083	3PEST_Rouge;
227	IBMR=Rouge	IBMR=Rouge	2	2	0,421	0,167	0,960	1,701	1,768	1MINE_Vert;1MOOX_Jaune;
224	IBMR=Rouge	IBMR=Rouge	2	1	0,474	0,167	0,960	1,658	1,734	1MOOX_Jaune;1PHOS_Jaune;
215	IBMR=Rouge	IBMR=Rouge	1	1	0,421	0,083	0,960	1,303	1,337	1 EPRV_Vert
228	IBMR=Rouge	IBMR=Rouge	2	2	0,421	0,167	0,960	1,243	1,311	1MOOX_Jaune;1NITR_Jaune;
229	IBMR=Rouge	IBMR=Rouge	3	3	0,421	0,250	0,960	1,170	1,271	1MOOX_Jaune;2NITR_Jaune;
247	IPR=Bleu	IPR=Bleu	3	3	0,479	0,250	0,960	1,353	1,468	3MOOX_Bleu;
237	IPR=Bleu	IPR=Bleu	1	1	0,404	0,083	0,960	1,367	1,399	1AZOT_Bleu;
245	IPR=Bleu	IPR=Bleu	2	2	0,573	0,167	0,800	1,249	1,325	2MOOX_Bleu;
244	IPR=Bleu	IPR=Bleu	2	2	0,406	0,167	0,960	1,236	1,301	1MOOX_Bleu;1PHOS_Bleu;
241	IPR=Bleu	IPR=Bleu	1	1	0,657	0,083	0,520	1,228	1,256	1MOOX_Bleu;
277	IPR=Vert	IPR=Bleu	2	2	0,409	0,167	0,800		0,055	2MOOX_Bleu;
278	IPR=Vert	IBD=Bleu	2	2	0,469	0,167	0,640		0,050	2PHOS_Bleu;
276	IPR=Vert	IBD=Bleu	1	1	0,576	0,083	0,520		0,025	1PHOS_Bleu;
274	IPR=Vert	IPR=Bleu	1	1	0,502	0,083	0,520		0,022	1MOOX_Bleu;
273	IPR=Vert	IPR=Rouge	1	1	0,424	0,083	0,320		0,011	1PEST_Rouge;
255	IPR=Jaune	IBGN=Jaune	2	2	0,418	0,167	0,360		0,025	2PEST_Rouge;
250	IPR=Jaune	IBD=Bleu	1	1	0,483	0,083	0,520		0,021	1PHOS_Bleu;
254	IPR=Jaune	IPR=Bleu	1	1	0,423	0,083	0,520		0,018	1MOOX_Bleu;
253	IPR=Jaune	IPR=Rouge	1	1	0,471	0,083	0,320		0,013	1PEST_Rouge;
252	IPR=Jaune	IBMR=Orange	1	1	0,461	0,083	0,160		0,006	1NITR_Jaune;
262	IPR=Orange	IBMR=Orange	2	2	0,418	0,167	0,520		0,036	2NITR_Jaune;
261	IPR=Orange	IPR=Rouge	1	1	0,404	0,083	0,880		0,030	1MPOR_Orange;
263	IPR=Orange	IBGN=Jaune	2	2	0,457	0,167	0,360		0,027	2PEST_Rouge;
259	IPR=Orange	IBMR=Rouge	1	1	0,440	0,083	0,520		0,019	1MOOX_Jaune;
258	IPR=Orange	IPR=Rouge	1	1	0,514	0,083	0,320		0,014	1PEST_Rouge;
269	IPR=Rouge	IPR=Rouge	1	1	0,441	0,083	0,880	1,048	1,080	1MPOR_Orange;
267	IPR=Rouge	IPR=Rouge	1	1	0,541	0,083	0,320	1,005	1,020	1PEST_Rouge;
270	IPR=Rouge	IBGN=Jaune	2	2	0,476	0,167	0,360		0,029	2PEST_Rouge;
266	IPR=Rouge	IBMR=Rouge	1	1	0,429	0,083	0,520		0,019	1MOOX_Jaune;
268	IPR=Rouge	IBMR=Orange	1	1	0,482	0,083	0,160		0,006	1NITR_Jaune;

ANNEXE 4 c: Caractéristiques des cinq motifs sélectionnés par contexte, sur la base de leur combinaison P la plus élevée, extraits pour la longueur de séquences 12 mois, avec le SEQ-eau

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E	Items du motif
51	IBD=Bleu	IBD=Bleu	5	5	0,545	0,625	0,960	1,3989	1,726	5PHOS_Bleu;
47	IBD=Bleu	IBD=Bleu	4	3	0,539	0,500	0,960	1,428	1,687	1ACID_vert;3PHOS_Bleu;
50	IBD=Bleu	IBD=Bleu	4	4	0,644	0,500	0,960	1,345	1,654	4PHOS_Bleu;
48	IBD=Bleu	IBD=Bleu	4	4	0,535	0,500	0,960	1,3861	1,643	1ACID_vert;3PHOS_Bleu;
46	IBD=Bleu	IBD=Bleu	3	3	0,546	0,375	0,960	1,3812	1,578	1ACID_vert;2PHOS_Bleu;
80	IBD=Vert	IBMR=Orange	2	2	0,541	0,250	0,480		0,065	2PEST_Rouge;
79	IBD=Vert	IPR=Bleu	1	1	0,573	0,125	0,720		0,052	1MOOX_Bleu;
75	IBD=Vert	IBD=Bleu	1	1	0,559	0,125	0,560		0,039	1PHOS_Bleu;
78	IBD=Vert	IPR=Rouge	1	1	0,568	0,125	0,360		0,026	1PEST_Rouge;
77	IBD=Vert	IBMR=Orange	1	1	0,617	0,125	0,320		0,025	1NITR_Jaune;
58	IBD=Jaune	IBMR=Orange	2	2	0,590	0,250	0,640		0,094	2NITR_Jaune;
57	IBD=Jaune	IBMR=Orange	2	2	0,566	0,250	0,480		0,068	2PEST_Rouge;
53	IBD=Jaune	IBMR=Rouge	1	1	0,585	0,125	0,560		0,041	1MOOX_Jaune;
54	IBD=Jaune	IBMR=Orange	1	1	0,710	0,125	0,320		0,028	1NITR_Jaune;
55	IBD=Jaune	IPR=Rouge	1	1	0,596	0,125	0,360		0,027	1PEST_Rouge;
65	IBD=Orange	IBMR=Orange	2	2	0,598	0,250	0,640		0,096	2NITR_Jaune;
63	IBD=Orange	IBMR=Rouge	1	1	0,608	0,125	0,880		0,067	1MINE_Vert;
66	IBD=Orange	IBMR=Orange	2	2	0,537	0,250	0,480		0,064	2PEST_Rouge;
62	IBD=Orange	IBMR=Rouge	1	1	0,559	0,125	0,560		0,039	1MOOX_Jaune;
64	IBD=Orange	IBMR=Orange	1	1	0,730	0,125	0,320		0,029	1NITR_Jaune;
72	IBD=Rouge	IBGN=Rouge	2	2	0,549	0,250	0,920		0,126	2PEST_Orange;
73	IBD=Rouge	IBMR=Orange	2	2	0,585	0,250	0,480		0,070	2PEST_Rouge;
69	IBD=Rouge	IBMR=Rouge	1	1	0,561	0,125	0,880		0,062	1MINE_Vert;
71	IBD=Rouge	IPR=Rouge	1	1	0,610	0,125	0,360		0,027	1PEST_Rouge;
68	IBD=Rouge	IBMR=Orange	1	1	0,598	0,125	0,320		0,024	1NITR_Jaune;
85	IBGN=Bleu	IBD=Bleu	2	2	0,587	0,250	0,720		0,106	2PHOS_Bleu;
82	IBGN=Bleu	IPR=Bleu	1	1	0,572	0,125	0,720		0,051	1MOOX_Bleu;
84	IBGN=Bleu	IBD=Bleu	1	1	0,675	0,125	0,560		0,047	1PHOS_Bleu;
83	IBGN=Bleu	IBMR=Rouge	1	1	0,595	0,125	0,000		0,000	1ACID_vert;
143	IBGN=Vert	IBD=Bleu	1	1	0,580	0,125	0,560		0,041	1PHOS_Bleu;
144	IBGN=Vert	IPR=Rouge	1	1	0,535	0,125	0,360		0,024	1PEST_Rouge;
145	IBGN=Vert	IBMR=Rouge	1	1	0,615	0,125	0,000		0,000	1ACID_vert;
92	IBGN=Jaune	IBMR=Orange	3	3	0,532	0,375	0,840		0,168	3PEST_Rouge;
91	IBGN=Jaune	IBMR=Orange	2	2	0,605	0,250	0,480		0,073	2PEST_Rouge;
89	IBGN=Jaune	IBMR=Rouge	1	1	0,549	0,125	0,560		0,038	1MOOX_Jaune;
88	IBGN=Jaune	IPR=Rouge	1	1	0,635	0,125	0,360		0,029	1PEST_Rouge;
90	IBGN=Jaune	IBMR=Orange	1	1	0,648	0,125	0,320		0,026	1NITR_Jaune;
96	IBGN=Orange	IBGN=Rouge	1	1	0,534	0,125	0,960		0,064	1MPOR_Jaune;
95	IBGN=Orange	IBMR=Rouge	1	1	0,530	0,125	0,840		0,056	1PHOS_Jaune;
97	IBGN=Orange	IBMR=Rouge	1	1	0,583	0,125	0,560		0,041	1MOOX_Jaune;
94	IBGN=Orange	IBMR=Orange	1	1	0,656	0,125	0,320		0,026	1NITR_Jaune;
98	IBGN=Orange	IPR=Rouge	1	1	0,563	0,125	0,360		0,025	1PEST_Rouge;
120	IBGN=Rouge	IBGN=Rouge	7	7	0,533	0,875	0,960	3,2	3,648	2NITR_Jaune;3MPOR_Jaune;2MPMI_Orange;
124	IBGN=Rouge	IBGN=Rouge	7	7	0,533	0,875	0,960	2,7178	3,166	4MPOR_Jaune;3MPMI_Orange;
132	IBGN=Rouge	IBGN=Rouge	7	7	0,533	0,875	0,960	2,6811	3,129	4MPOR_Jaune;3MPMI_Orange;
135	IBGN=Rouge	IBGN=Rouge	6	4	0,533	0,750	0,960	2,6811	3,065	4MPOR_Jaune;2MPMI_Orange;
114	IBGN=Rouge	IBGN=Rouge	5	5	0,567	0,625	0,960	2,7026	3,043	3MPOR_Jaune;2MPMI_Orange;
6	I2M2=Bleu	IBD=Bleu	3	3	0,579	0,375	0,880		0,191	3PHOS_Bleu;
5	I2M2=Bleu	IBD=Bleu	2	2	0,684	0,250	0,720		0,123	2PHOS_Bleu;
1	I2M2=Bleu	IBGN=Rouge	1	1	0,541	0,125	0,920		0,062	1PEST_Orange;
4	I2M2=Bleu	IBD=Bleu	1	1	0,769	0,125	0,560		0,054	1PHOS_Bleu;
2	I2M2=Bleu	IPR=Bleu	1	1	0,596	0,125	0,720		0,054	1MOOX_Bleu;
35	I2M2=Vert	IBD=Bleu	2	2	0,549	0,250	0,720		0,099	2PHOS_Bleu;
34	I2M2=Vert	IBD=Bleu	1	1	0,657	0,125	0,560		0,046	1PHOS_Bleu;
33	I2M2=Vert	IPR=Rouge	1	1	0,534	0,125	0,360		0,024	1PEST_Rouge;
32	I2M2=Vert	IBMR=Rouge	1	1	0,630	0,125	0,000		0,000	1ACID_vert;
11	I2M2=Jaune	IBMR=Orange	2	2	0,535	0,250	0,640		0,086	2NITR_Jaune;
12	I2M2=Jaune	IBMR=Orange	2	2	0,547	0,250	0,480		0,066	2PEST_Rouge;
8	I2M2=Jaune	IBMR=Orange	1	1	0,644	0,125	0,320		0,026	1NITR_Jaune;
10	I2M2=Jaune	IPR=Rouge	1	1	0,571	0,125	0,360		0,026	1PEST_Rouge;
9	I2M2=Jaune	IBMR=Rouge	1	1	0,578	0,125	0,000		0,000	1ACID_vert;
21	I2M2=Orange	IBMR=Orange	3	3	0,543	0,375	0,840		0,171	3PEST_Rouge;
19	I2M2=Orange	IBMR=Orange	2	2	0,622	0,250	0,640		0,100	2NITR_Jaune;

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E	Items du motif
20	I2M2=Orange	IBMR=Orange	2	2	0,607	0,250	0,480		0,073	2PEST_Rouge;
16	I2M2=Orange	IBMR=Rouge	1	1	0,540	0,125	0,840		0,057	1PHOS_Jaune;
18	I2M2=Orange	IBMR=Rouge	1	1	0,629	0,125	0,560		0,044	1MOOX_Jaune;
23	I2M2=Rouge	I2M2=Rouge	1	1	0,616	0,125	0,920	1,3906	1,461	1AZOT_Jaune;
29	I2M2=Rouge	IBMR=Orange	2	2	0,573	0,250	0,640		0,092	2NITR_Jaune;
30	I2M2=Rouge	IBMR=Orange	2	2	0,575	0,250	0,480		0,069	2PEST_Rouge;
27	I2M2=Rouge	IBMR=Rouge	1	1	0,608	0,125	0,840		0,064	1PHOS_Jaune;
25	I2M2=Rouge	IBMR=Rouge	1	1	0,656	0,125	0,560		0,046	1MOOX_Jaune;
154	IBMR=Bleu	IBD=Bleu	3	3	0,579	0,375	0,880		0,191	3PHOS_Bleu;
153	IBMR=Bleu	IBGN=Rouge	2	2	0,565	0,250	0,920		0,130	2PEST_Orange;
151	IBMR=Bleu	IPR=Bleu	2	2	0,561	0,250	0,920		0,129	2MOOX_Bleu;
152	IBMR=Bleu	IBD=Bleu	2	2	0,647	0,250	0,720		0,116	2PHOS_Bleu;
147	IBMR=Bleu	IBGN=Rouge	1	1	0,581	0,125	0,960		0,070	1PEST_Orange;
201	IBMR=Vert	IPR=Bleu	1	1	0,585	0,125	0,720		0,053	1MOOX_Bleu;
203	IBMR=Vert	IBD=Bleu	1	1	0,564	0,125	0,560		0,039	1PHOS_Bleu;
202	IBMR=Vert	IBMR=Orange	1	1	0,539	0,125	0,320		0,022	1NITR_Jaune;
204	IBMR=Vert	IBMR=Rouge	1	1	0,587	0,125	0,000		0,000	1ACID_vert;
161	IBMR=Jaune	IBMR=Orange	2	2	0,601	0,250	0,640		0,096	2NITR_Jaune;
162	IBMR=Jaune	IBMR=Orange	2	2	0,581	0,250	0,480		0,070	2PEST_Rouge;
157	IBMR=Jaune	IBMR=Rouge	1	1	0,603	0,125	0,560		0,042	1MOOX_Jaune;
158	IBMR=Jaune	IBMR=Orange	1	1	0,702	0,125	0,320		0,028	1NITR_Jaune;
174	IBMR=Orange	IBMR=Orange	3	3	0,646	0,375	0,920	1,2885	1,511	3NITR_Jaune;
179	IBMR=Orange	IBMR=Orange	4	4	0,536	0,500	0,960	1,116	1,373	4PEST_Rouge;
171	IBMR=Orange	IBMR=Orange	2	2	0,740	0,250	0,640	1,1891	1,307	2NITR_Jaune;
175	IBMR=Orange	IBMR=Orange	2	2	0,531	0,250	0,960	1,1434	1,271	1NITR_Jaune;1PEST_Rouge;
177	IBMR=Orange	IBMR=Orange	3	3	0,583	0,375	0,840	1,0749	1,259	3PEST_Rouge;
191	IBMR=Rouge	IBMR=Rouge	3	3	0,571	0,375	0,960	1,8596	2,065	1MINE_Vert;2MOOX_Jaune;
199	IBMR=Rouge	IBMR=Rouge	3	3	0,571	0,375	0,960	1,6135	1,819	3MOOX_Jaune;
197	IBMR=Rouge	IBMR=Rouge	6	5	0,571	0,750	0,960	1,3888	1,800	3MOOX_Jaune;3NITR_Jaune;
190	IBMR=Rouge	IBMR=Rouge	2	2	0,571	0,250	0,960	1,5584	1,696	1PAES_Jaune;1MOOX_Jaune;
182	IBMR=Rouge	IBMR=Rouge	1	1	0,714	0,125	0,960	1,4286	1,514	1PAES_Jaune;
211	IPR=Bleu	IPR=Bleu	3	3	0,557	0,375	0,960	1,1874	1,388	3MOOX_Bleu;
210	IPR=Bleu	IPR=Bleu	2	2	0,646	0,250	0,920	1,1518	1,300	2MOOX_Bleu;
208	IPR=Bleu	IPR=Bleu	1	1	0,710	0,125	0,720	1,156	1,220	1MOOX_Bleu;
209	IPR=Bleu	IBD=Bleu	2	2	0,593	0,250	0,720		0,107	2PHOS_Bleu;
207	IPR=Bleu	IBD=Bleu	1	1	0,710	0,125	0,560		0,050	1PHOS_Bleu;
236	IPR=Vert	IBD=Bleu	2	2	0,531	0,250	0,720		0,096	2PHOS_Bleu;
235	IPR=Vert	IPR=Bleu	1	1	0,594	0,125	0,720		0,053	1MOOX_Bleu;
234	IPR=Vert	IBD=Bleu	1	1	0,625	0,125	0,560		0,044	1PHOS_Bleu;
233	IPR=Vert	IBMR=Rouge	1	1	0,651	0,125	0,000		0,000	1ACID_vert;
216	IPR=Jaune	IBD=Bleu	1	1	0,560		0,038			1PHOS_Bleu;
215	IPR=Jaune	IPR=Rouge	1	1	0,360		0,025			1PEST_Rouge;
214	IPR=Jaune	IBMR=Orange	1	1	0,320		0,022			1NITR_Jaune;
213	IPR=Jaune	IBMR=Rouge	1	1	0,000		0,000			1ACID_vert;
223	IPR=Orange	IBMR=Orange	2	2	0,640		0,088			2NITR_Jaune;
222	IPR=Orange	IBMR=Orange	2	2	0,480		0,068			2PEST_Rouge;
218	IPR=Orange	IBMR=Rouge	1	1	0,560		0,038			1MOOX_Jaune;
221	IPR=Orange	IPR=Rouge	1	1	0,360		0,029			1PEST_Rouge;
220	IPR=Orange	IBMR=Orange	1	1	0,320		0,027			1NITR_Jaune;
229	IPR=Rouge	IPR=Rouge	1	1	0,360	1,0063	1,036			1PEST_Rouge;
231	IPR=Rouge	IBMR=Orange	3	3	0,840		0,173			3PEST_Rouge;
230	IPR=Rouge	IBMR=Orange	2	2	0,480		0,073			2PEST_Rouge;
225	IPR=Rouge	IBMR=Rouge	1	1	0,880		0,062			1MPOR_Orange
227	IPR=Rouge	IBMR=Rouge	1	1	0,560		0,039			1MOOX_Jaune;

ANNEXE 4 d: Caractéristiques des cinq motifs sélectionnés par contexte, sur la base de leur combinaison P la plus élevée, extraits pour la longueur de séquences 18 mois, avec le SEQ-eau

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E		Items du motif
53	IBD=Bleu	IBD=Bleu	6	6	0,637	1,000	0,960	1,379	1,990		6PHOS_Bleu;
50	IBD=Bleu	IBD=Bleu	5	5	0,715	0,833	0,960	1,340	1,912		5PHOS_Bleu;
52	IBD=Bleu	IBD=Bleu	5	5	0,603	0,833	0,960	1,380	1,863		1ACID_vert;4PHOS_Bleu;
48	IBD=Bleu	IBD=Bleu	4	4	0,784	0,667	0,920	1,290	1,771		4PHOS_Bleu;
47	IBD=Bleu	IBD=Bleu	4	4	0,616	0,667	0,960	1,356	1,750		1ACID_vert;3PHOS_Bleu;
72	IBD=Vert	IBD=Bleu	1	1	0,612	0,167	0,720		0,073		1PHOS_Bleu;
73	IBD=Vert	IPR=Bleu	1	1	0,610	0,167	0,720		0,073		1MOOX_Bleu;
74	IBD=Vert	IBMR=Orange	1	1	0,650	0,167	0,360		0,039		1NITR_Jaune;
75	IBD=Vert	IBMR=Rouge	1	1	0,720	0,167	0,000		0,000		1ACID_vert;
59	IBD=Jaune	IBMR=Orange	2	2	0,660	0,333	0,640		0,141		2NITR_Jaune;
58	IBD=Jaune	IPR=Rouge	1	1	0,613	0,167	0,600		0,061		1PEST_Rouge;
57	IBD=Jaune	IBMR=Rouge	1	1	0,649	0,167	0,520		0,056		1MOOX_Jaune;
56	IBD=Jaune	IBMR=Orange	1	1	0,758	0,167	0,360		0,045		1NITR_Jaune;
55	IBD=Jaune	IBMR=Rouge	1	1	0,653	0,167	0,000		0,000		1ACID_vert;
61	IBD=Orange	IBD=Orange	1	1	0,687	0,167	0,880	1,247	1,348		1MINE_Vert;
65	IBD=Orange	IBMR=Orange	2	2	0,673	0,333	0,640		0,144		2NITR_Jaune;
62	IBD=Orange	IBMR=Rouge	1	1	0,638	0,167	0,520		0,055		1MOOX_Jaune;
64	IBD=Orange	IBMR=Orange	1	1	0,764	0,167	0,360		0,046		1NITR_Jaune;
63	IBD=Orange	IBMR=Rouge	1	1	0,661	0,167	0,000		0,000		1ACID_vert;
67	IBD=Rouge	IBD=Orange	1	1	0,622	0,167	0,880		0,091		1MINE_Vert;
68	IBD=Rouge	IPR=Rouge	1	1	0,610	0,167	0,600		0,061		1PEST_Rouge;
70	IBD=Rouge	IBMR=Orange	1	1	0,671	0,167	0,360		0,040		1NITR_Jaune;
69	IBD=Rouge	IBMR=Rouge	1	1	0,646	0,167	0,000		0,000		1ACID_vert;
80	IBGN=Bleu	IBD=Bleu	2	2	0,650	0,333	0,760		0,165		2PHOS_Bleu;
79	IBGN=Bleu	IBD=Bleu	1	1	0,714	0,167	0,720		0,086		1PHOS_Bleu;
77	IBGN=Bleu	IPR=Bleu	1	1	0,605	0,167	0,720		0,073		1MOOX_Bleu;
78	IBGN=Bleu	IBMR=Rouge	1	1	0,663	0,167	0,000		0,000		1ACID_vert;
116	IBGN=Vert	IBD=Bleu	1	1	0,621	0,167	0,720		0,075		1PHOS_Bleu;
117	IBGN=Vert	IBMR=Rouge	1	1	0,704	0,167	0,000		0,000		1ACID_vert;
87	IBGN=Jaune	IBMR=Orange	3	3	0,601	0,500	0,840		0,252		3PEST_Rouge;
86	IBGN=Jaune	IPR=Rouge	2	2	0,623	0,333	0,760		0,158		2PEST_Rouge;
85	IBGN=Jaune	IPR=Rouge	1	1	0,661	0,167	0,600		0,066		1PEST_Rouge;
82	IBGN=Jaune	IBMR=Rouge	1	1	0,608	0,167	0,520		0,053		1MOOX_Jaune;
83	IBGN=Jaune	IBMR=Orange	1	1	0,701	0,167	0,360		0,042		1NITR_Jaune;
90	IBGN=Orange	IBMR=Rouge	1	1	0,665	0,167	0,520		0,058		1MOOX_Jaune;
89	IBGN=Orange	IBMR=Orange	1	1	0,710	0,167	0,360		0,043		1NITR_Jaune;
91	IBGN=Orange	IBMR=Rouge	1	1	0,641	0,167	0,000		0,000		1ACID_vert;
104	IBGN=Rouge	IBGN=Rouge	4	4	0,600	0,667	0,960	2,117	2,501		1NITR_Jaune;
113	IBGN=Rouge	IBGN=Rouge	6	6	0,633	1,000	0,960	1,880	2,488		1NITR_Jaune;5MPOR_Jaune
114	IBGN=Rouge	IBGN=Rouge	6	6	0,600	1,000	0,960	1,839	2,415		1NITR_Jaune;5MPOR_Jaune
111	IBGN=Rouge	IBGN=Rouge	6	6	0,667	1,000	0,960	1,744	2,384		6MPOR_Jaune
108	IBGN=Rouge	IBGN=Rouge	5	5	0,733	0,833	0,960	1,700	2,286		5MPOR_Jaune
6	I2M2=Bleu	IBD=Bleu	4	4	0,608	0,667	0,920		0,373		4PHOS_Bleu;
5	I2M2=Bleu	IBD=Bleu	3	3	0,683	0,500	0,840		0,287		3PHOS_Bleu;
4	I2M2=Bleu	IBD=Bleu	2	2	0,750	0,333	0,760		0,190		2PHOS_Bleu;
1	I2M2=Bleu	IBD=Bleu	1	1	0,800	0,167	0,880		0,117		1PHOS_Bleu;
2	I2M2=Bleu	IPR=Bleu	1	1	0,618	0,167	0,720		0,074		1MOOX_Bleu;
32	I2M2=Vert	IBD=Bleu	2	2	0,617	0,333	0,760		0,156		2PHOS_Bleu;
31	I2M2=Vert	IBD=Bleu	1	1	0,705	0,167	0,720		0,085		1PHOS_Bleu;
30	I2M2=Vert	IBMR=Rouge	1	1	0,712	0,167	0,000		0,000		1ACID_vert;
9	I2M2=Jaune	IBMR=Orange	1	1	0,690	0,167	0,360		0,041		1NITR_Jaune;
8	I2M2=Jaune	IBMR=Rouge	1	1	0,669	0,167	0,000		0,000		1ACID_vert;
17	I2M2=Orange	IBMR=Orange	3	3	0,603	0,500	0,920		0,277		3NITR_Jaune;
16	I2M2=Orange	IPR=Rouge	2	2	0,621	0,333	0,760		0,157		2PEST_Rouge;
15	I2M2=Orange	IBMR=Orange	2	2	0,685	0,333	0,640		0,146		2NITR_Jaune;
14	I2M2=Orange	IPR=Rouge	1	1	0,663	0,167	0,600		0,066		1PEST_Rouge;
12	I2M2=Orange	IBMR=Rouge	1	1	0,697	0,167	0,520		0,060		1MOOX_Jaune;
20	I2M2=Rouge	I2M2=Rouge	1	1	0,647	0,167	0,920	1,080	1,179		1PHOS_Jaune;
28	I2M2=Rouge	IBGN=Rouge	2	2	0,610	0,333	0,840		0,171		1MOOX_Jaune;1NITR_Jaune;
26	I2M2=Rouge	IPR=Rouge	2	2	0,602	0,333	0,760		0,152		2PEST_Rouge;
27	I2M2=Rouge	IBMR=Orange	2	2	0,666	0,333	0,640		0,142		2NITR_Jaune;
123	IBMR=Bleu	IBD=Bleu	3	3	0,631	0,500	0,840		0,265		3PHOS_Bleu;
122	IBMR=Bleu	IBD=Bleu	2	2	0,677	0,333	0,760		0,171		2PHOS_Bleu;

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E		Items du motif
120	IBMR=Bleu	IBD=Bleu	1	1	0,732	0,167	0,720		0,088		1PHOS_Bleu;
121	IBMR=Bleu	IPR=Bleu	1	1	0,641	0,167	0,720		0,077		1MOOX_Bleu;
119	IBMR=Bleu	IBMR=Rouge	1	1	0,631	0,167	0,000		0,000		1ACID_vert;
169	IBMR=Vert	IPR=Bleu	1	1	0,619	0,167	0,720		0,074		1MOOX_Bleu;
170	IBMR=Vert	IBD=Bleu	1	1	0,613	0,167	0,720		0,074		1PHOS_Bleu;
171	IBMR=Vert	IBMR=Rouge	1	1	0,658	0,167	0,000		0,000		1ACID_vert;
129	IBMR=Jaune	IBMR=Orange	2	2	0,651	0,333	0,640		0,139		2NITR_Jaune;
126	IBMR=Jaune	IPR=Rouge	1	1	0,612	0,167	0,600		0,061		1PEST_Rouge;
125	IBMR=Jaune	IBMR=Rouge	1	1	0,654	0,167	0,520		0,057		1MOOX_Jaune;
128	IBMR=Jaune	IBMR=Orange	1	1	0,724	0,167	0,360		0,043		1NITR_Jaune;
127	IBMR=Jaune	IBMR=Rouge	1	1	0,671	0,167	0,000		0,000		1ACID_vert;
144	IBMR=Orange	IBMR=Orange	4	4	0,630	0,667	0,960	1,179	1,582		4NITR_Jaune;
141	IBMR=Orange	IBMR=Orange	3	3	0,724	0,500	0,920	1,201	1,534		3NITR_Jaune;
140	IBMR=Orange	IBMR=Orange	3	3	0,604	0,500	0,960	1,243	1,533		1MOOX_Jaune;2NITR_Jaune;
143	IBMR=Orange	IBMR=Orange	3	3	0,641	0,500	0,960	1,223	1,531		1MOOX_Jaune;2NITR_Jaune;
145	IBMR=Orange	IBMR=Orange	4	4	0,604	0,667	0,960	1,049	1,435		4PEST_Rouge;
156	IBMR=Rouge	IBMR=Rouge	2	2	0,667	0,333	0,960	1,818	2,032		1PAES_Jaune;1MOOX_Jaune;
160	IBMR=Rouge	IBMR=Rouge	2	2	0,619	0,333	0,960	1,467	1,665		1ACID_vert;1MOOX_Jaune;
163	IBMR=Rouge	IBMR=Rouge	3	3	0,619	0,500	0,960	1,354	1,651		2MOOX_Jaune;1AZOT_Jaune;
165	IBMR=Rouge	IBMR=Rouge	2	2	0,714	0,333	0,960	1,236	1,464		2MOOX_Jaune;
153	IBMR=Rouge	IBMR=Rouge	1	1	0,714	0,167	0,960	1,339	1,454		1PAES_Rouge;
178	IPR=Bleu	IPR=Bleu	3	3	0,648	0,500	0,960	1,234	1,545		3MOOX_Bleu;
177	IPR=Bleu	IPR=Bleu	2	2	0,705	0,333	0,960	1,209	1,434		2MOOX_Bleu;
174	IPR=Bleu	IPR=Bleu	1	1	0,743	0,167	0,720	1,159	1,248		1MOOX_Bleu;
179	IPR=Bleu	IBD=Bleu	3	3	0,603	0,500	0,840		0,253		3PHOS_Bleu;
176	IPR=Bleu	IBD=Bleu	2	2	0,669	0,333	0,760		0,170		2PHOS_Bleu;
200	IPR=Vert	IBD=Bleu	1	1	0,672	0,167	0,720		0,081		1PHOS_Bleu;
198	IPR=Vert	IPR=Bleu	1	1	0,631	0,167	0,720		0,076		1MOOX_Bleu;
199	IPR=Vert	IBMR=Rouge	1	1	0,727	0,167	0,000		0,000		1ACID_vert;
181	IPR=Jaune	IBMR=Orange	1	1	0,606	0,167	0,360		0,036		1NITR_Jaune;
182	IPR=Jaune	IBMR=Rouge	1	1	0,693	0,167	0,000		0,000		1ACID_vert;
188	IPR=Orange	IPR=Rouge	2	2	0,617	0,333	0,760		0,156		2PEST_Rouge;
189	IPR=Orange	IBMR=Orange	2	2	0,613	0,333	0,640		0,131		2NITR_Jaune;
186	IPR=Orange	IPR=Rouge	1	1	0,671	0,167	0,600		0,067		1PEST_Rouge;
185	IPR=Orange	IBMR=Rouge	1	1	0,607	0,167	0,520		0,053		1MOOX_Jaune;
187	IPR=Orange	IBMR=Orange	1	1	0,690	0,167	0,360		0,041		1NITR_Jaune;
195	IPR=Rouge	IPR=Rouge	2	2	0,643	0,333	0,760	1,003	1,166		2PEST_Rouge;
193	IPR=Rouge	IPR=Rouge	1	1	0,706	0,167	0,600	1,053	1,124		1PEST_Rouge;
196	IPR=Rouge	IBMR=Orange	3	3	0,619	0,500	0,840		0,260		3PEST_Rouge;
194	IPR=Rouge	IBMR=Rouge	1	1	0,606	0,167	0,520		0,053		1MOOX_Jaune;
191	IPR=Rouge	IBMR=Orange	1	1	0,672	0,167	0,360		0,040		1NITR_Jaune;

ANNEXE 4 e: Caractéristiques des cinq motifs sélectionnés par contexte, sur la base de leur combinaison P la plus élevée, extraits pour la longueur de séquences 24 mois, avec le SEQ-eau

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E		Items du motif
58	IBD=Bleu	IBD=Bleu	7	7	0,669	0,583	0,960		1,34089	1,715	7PHOS_Bleu;
57	IBD=Bleu	IBD=Bleu	6	6	0,718	0,500	0,960		1,32913	1,674	6PHOS_Bleu;
54	IBD=Bleu	IBD=Bleu	5	5	0,751	0,417	0,960		1,29076	1,591	5PHOS_Bleu;
55	IBD=Bleu	IBD=Bleu	5	5	0,651	0,417	0,960		1,31131	1,572	1ACID_vert;4PHOS_Bleu;
51	IBD=Bleu	IBD=Bleu	4	4	0,809	0,333	0,960		1,26591	1,525	4PHOS_Bleu;
83	IBD=Vert	IBMR=Rouge	2	2	0,654	0,167	0,480			0,052	2PEST_Rouge;
79	IBD=Vert	IPR=Bleu	1	1	0,657	0,083	0,720			0,039	1MOOX_Bleu;
81	IBD=Vert	IBMR=Rouge	1	1	0,673	0,083	0,440			0,025	1PEST_Rouge;
80	IBD=Vert	IBMR=Orange	1	1	0,680	0,083	0,400			0,023	1NITR_Jaune;
82	IBD=Vert	IPR=Rouge	1	1	0,756	0,083	0,040			0,003	1ACID_vert;
66	IBD=Jaune	IBMR=Rouge	3	3	0,656	0,250	0,680			0,112	3PEST_Rouge;
64	IBD=Jaune	IBMR=Orange	2	2	0,696	0,167	0,760			0,088	2NITR_Jaune;
65	IBD=Jaune	IBMR=Rouge	2	2	0,686	0,167	0,480			0,055	2PEST_Rouge;
61	IBD=Jaune	IBGN=Rouge	1	1	0,701	0,083	0,520			0,030	1MOOX_Jaune;
62	IBD=Jaune	IBMR=Orange	1	1	0,782	0,083	0,400			0,026	1NITR_Jaune;
72	IBD=Orange	IBMR=Orange	2	2	0,706	0,167	0,760			0,089	2NITR_Jaune;
68	IBD=Orange	IBMR=Rouge	1	1	0,715	0,083	0,920			0,055	1MINE_Vert;
70	IBD=Orange	IBGN=Rouge	1	1	0,674	0,083	0,520			0,029	1MOOX_Jaune;
71	IBD=Orange	IBMR=Orange	1	1	0,798	0,083	0,400			0,027	1NITR_Jaune;
69	IBD=Orange	IPR=Rouge	1	1	0,686	0,083	0,040			0,002	1ACID_vert;
77	IBD=Rouge	IBMR=Rouge	2	2	0,720	0,167	0,480			0,058	2PEST_Rouge;
75	IBD=Rouge	IBMR=Rouge	1	1	0,744	0,083	0,440			0,027	1PEST_Rouge;
74	IBD=Rouge	IBMR=Orange	1	1	0,707	0,083	0,400			0,024	1NITR_Jaune;
76	IBD=Rouge	IPR=Rouge	1	1	0,671	0,083	0,040			0,002	1ACID_vert;
88	IBGN=Bleu	IBD=Bleu	2	2	0,673	0,167	0,800			0,090	2PHOS_Bleu;
86	IBGN=Bleu	IBD=Bleu	1	1	0,738	0,083	0,720			0,044	1PHOS_Bleu;
87	IBGN=Bleu	IPR=Bleu	1	1	0,660	0,083	0,720			0,040	1MOOX_Bleu;
85	IBGN=Bleu	IPR=Rouge	1	1	0,705	0,083	0,040			0,002	1ACID_vert;
116	IBGN=Vert	IPR=Rouge	1	1	0,738	0,083	0,040			0,002	1ACID_vert;
95	IBGN=Jaune	IBMR=Rouge	3	3	0,671	0,250	0,680			0,114	3PEST_Rouge;
94	IBGN=Jaune	IBMR=Rouge	2	2	0,693	0,167	0,480			0,055	2PEST_Rouge;
90	IBGN=Jaune	IBGN=Rouge	1	1	0,675	0,083	0,520			0,029	1MOOX_Jaune;
93	IBGN=Jaune	IBMR=Rouge	1	1	0,727	0,083	0,440			0,027	1PEST_Rouge;
91	IBGN=Jaune	IBMR=Orange	1	1	0,725	0,083	0,400			0,024	1NITR_Jaune;
97	IBGN=Orange	IBGN=Rouge	1	1	0,722	0,083	0,520			0,031	1MOOX_Jaune;
98	IBGN=Orange	IBMR=Orange	1	1	0,742	0,083	0,400			0,025	1NITR_Jaune;
99	IBGN=Orange	IPR=Rouge	1	1	0,661	0,083	0,040			0,002	1ACID_vert;
112	IBGN=Rouge	IBGN=Rouge	7	7	0,667	0,583	0,960		2,27381	2,647	1MOOX_Jaune;1NITR_Jaune;5MPOR_Jaune
108	IBGN=Rouge	IBGN=Rouge	6	6	0,733	0,500	0,960		2,0598	2,412	1MOOX_Jaune;5MPOR_Jaune
113	IBGN=Rouge	IBGN=Rouge	7	7	0,667	0,583	0,960		2,00525	2,379	1MOOX_Jaune;6MPOR_Jaune
111	IBGN=Rouge	IBGN=Rouge	6	6	0,700	0,500	0,960		1,71881	2,055	6MPOR_Jaune
110	IBGN=Rouge	IBGN=Rouge	4	4	0,700	0,333	0,960		1,70318	1,927	1NITR_Jaune;3MPOR_Jaune
7	I2M2=Bleu	IBD=Bleu	3	3	0,713	0,250	0,880			0,157	3PHOS_Bleu;
6	I2M2=Bleu	IBD=Bleu	2	2	0,777	0,167	0,800			0,104	2PHOS_Bleu;
5	I2M2=Bleu	IBMR=Bleu	2	2	0,667	0,167	0,800			0,089	2PEST_Orange;
2	I2M2=Bleu	IBD=Bleu	1	1	0,831	0,083	0,720			0,050	1PHOS_Bleu;
3	I2M2=Bleu	IBMR=Bleu	1	1	0,672	0,083	0,760			0,043	1PEST_Orange;
35	I2M2=Vert	IBD=Bleu	1	1	0,720	0,083	0,720			0,043	1PHOS_Bleu;
37	I2M2=Vert	IBMR=Rouge	1	1	0,654	0,083	0,440			0,024	1PEST_Rouge;
36	I2M2=Vert	IPR=Rouge	1	1	0,741	0,083	0,040			0,002	1ACID_vert;
12	I2M2=Jaune	IBMR=Rouge	2	2	0,656	0,167	0,480			0,052	2PEST_Rouge;
11	I2M2=Jaune	IBMR=Rouge	1	1	0,679	0,083	0,440			0,025	1PEST_Rouge;
9	I2M2=Jaune	IBMR=Orange	1	1	0,715	0,083	0,400			0,024	1NITR_Jaune;
10	I2M2=Jaune	IPR=Rouge	1	1	0,705	0,083	0,040			0,002	1ACID_vert;
21	I2M2=Orange	IBMR=Orange	3	3	0,651	0,250	0,920			0,150	3NITR_Jaune;
22	I2M2=Orange	IBMR=Rouge	3	3	0,688	0,250	0,680			0,117	3PEST_Rouge;
20	I2M2=Orange	IBMR=Orange	2	2	0,728	0,167	0,760			0,092	2NITR_Jaune;
19	I2M2=Orange	IBMR=Rouge	2	2	0,710	0,167	0,480			0,057	2PEST_Rouge;
14	I2M2=Orange	IBMR=Rouge	1	1	0,655	0,083	0,880			0,048	1PHOS_Jaune;
24	I2M2=Rouge	I2M2=Rouge	1	1	0,725	0,083	0,960		1,22886	1,287	1AZOT_Jaune;
33	I2M2=Rouge	IBGN=Rouge	2	2	0,662	0,167	0,880			0,097	1MOOX_Jaune;1NITR_Jaune;
32	I2M2=Rouge	IBMR=Orange	2	2	0,702	0,167	0,760			0,089	2NITR_Jaune;
31	I2M2=Rouge	IBMR=Rouge	2	2	0,660	0,167	0,480			0,053	2PEST_Rouge;

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E		Items du motif
25	I2M2=Rouge	IBMR=Rouge	1	1	0,681	0,083	0,880		0,050		1PHOS_Jaune;
125	IBMR=Bleu	IBMR=Bleu	3	3	0,666	0,250	0,960	1,03661	1,196		3PEST_Orange;
122	IBMR=Bleu	IBMR=Bleu	2	2	0,707	0,167	0,800	1,06048	1,155		2PEST_Orange;
120	IBMR=Bleu	IBMR=Bleu	1	1	0,714	0,083	0,760	1,06266	1,108		1PEST_Orange;
124	IBMR=Bleu	IBD=Bleu	3	3	0,665	0,250	0,880		0,146		3PHOS_Bleu;
123	IBMR=Bleu	IBD=Bleu	2	2	0,707	0,167	0,800		0,094		2PHOS_Bleu;
245	IBMR=Vert	IBMR=Bleu	2	2	0,668	0,167	0,800		0,089		2PEST_Orange;
242	IBMR=Vert	IBMR=Bleu	1	1	0,686	0,083	0,760		0,043		1PEST_Orange;
243	IBMR=Vert	IPR=Bleu	1	1	0,667	0,083	0,720		0,040		1MOOX_Bleu;
244	IBMR=Vert	IBMR=Rouge	1	1	0,667	0,083	0,440		0,024		1PEST_Rouge;
241	IBMR=Vert	IPR=Rouge	1	1	0,700	0,083	0,040		0,002		1ACID_vert;
134	IBMR=Jaune	IBMR=Rouge	3	3	0,692	0,250	0,680		0,118		3PEST_Rouge;
132	IBMR=Jaune	IBMR=Orange	2	2	0,668	0,167	0,760		0,085		2NITR_Jaune;
133	IBMR=Jaune	IBMR=Rouge	2	2	0,719	0,167	0,480		0,058		2PEST_Rouge;
127	IBMR=Jaune	IBMR=Bleu	1	1	0,653	0,083	0,760		0,041		1PEST_Orange;
129	IBMR=Jaune	IBGN=Rouge	1	1	0,704	0,083	0,520		0,031		1MOOX_Jaune;
162	IBMR=Orange	IBMR=Orange	4	4	0,651	0,333	0,960	1,22645	1,435		1MOOX_Jaune;3NITR_Jaune;
164	IBMR=Orange	IBMR=Orange	4	4	0,703	0,333	0,960	1,19894	1,424		4NITR_Jaune;
160	IBMR=Orange	IBMR=Orange	3	3	0,677	0,250	0,960	1,22527	1,388		1MOOX_Jaune;2NITR_Jaune;
154	IBMR=Orange	IBMR=Orange	4	4	0,651	0,333	0,960	1,17814	1,386		2NITR_Jaune;2PEST_Rouge;
156	IBMR=Orange	IBMR=Orange	3	3	0,776	0,250	0,920	1,19291	1,371		3NITR_Jaune;
217	IBMR=Rouge	IBMR=Rouge	10	10	0,667	0,833	0,960	2,15365	2,687		2MOOX_Jaune;5PEST_Rouge;3MPOR_Orange;
211	IBMR=Rouge	IBMR=Rouge	7	7	0,667	0,583	0,960	2,29429	2,668		1PAES_Jaune;4PEST_Rouge;2MPOR_Orange;
203	IBMR=Rouge	IBMR=Rouge	5	5	0,714	0,417	0,960	2,36453	2,650		1PAES_Jaune;2PEST_Rouge;2MPOR_Orange;
199	IBMR=Rouge	IBMR=Rouge	5	5	0,667	0,417	0,960	2,24561	2,512		1PAES_Jaune;2MOOX_Jaune;2PEST_Rouge;
215	IBMR=Rouge	IBMR=Rouge	9	7	0,667	0,750	0,960	1,75342	2,233		1MOOX_Jaune;5PEST_Rouge;3MPOR_Orange;
254	IPR=Bleu	IPR=Bleu	3	3	0,683	0,250	0,960	1,22718	1,391		3MOOX_Bleu;
253	IPR=Bleu	IPR=Bleu	2	2	0,732	0,167	0,960	1,1738	1,291		2MOOX_Bleu;
249	IPR=Bleu	IPR=Bleu	1	1	0,779	0,083	0,720	1,12833	1,175		1MOOX_Bleu;
252	IPR=Bleu	IBD=Bleu	2	2	0,692	0,167	0,800		0,092		2PHOS_Bleu;
250	IPR=Bleu	IBD=Bleu	1	1	0,762	0,083	0,720		0,046		1PHOS_Bleu;
281	IPR=Vert	IBMR=Bleu	2	2	0,659	0,167	0,800		0,088		2PEST_Orange;
279	IPR=Vert	IBMR=Bleu	1	1	0,668	0,083	0,760		0,042		1PEST_Orange;
277	IPR=Vert	IBD=Bleu	1	1	0,697	0,083	0,720		0,042		1PHOS_Bleu;
280	IPR=Vert	IPR=Bleu	1	1	0,683	0,083	0,720		0,041		1MOOX_Bleu;
278	IPR=Vert	IPR=Rouge	1	1	0,762	0,083	0,040		0,003		1ACID_vert;
256	IPR=Jaune	IPR=Rouge	1	1	0,731	0,083	0,040		0,002		1ACID_vert;
263	IPR=Orange	IBMR=Rouge	3	3	0,655	0,250	0,680		0,111		3PEST_Rouge;
262	IPR=Orange	IBMR=Rouge	2	2	0,693	0,167	0,480		0,055		2PEST_Rouge;
258	IPR=Orange	IBGN=Rouge	1	1	0,665	0,083	0,520		0,029		1MOOX_Jaune;
261	IPR=Orange	IBMR=Rouge	1	1	0,736	0,083	0,440		0,027		1PEST_Rouge;
259	IPR=Orange	IBMR=Orange	1	1	0,721	0,083	0,400		0,024		1NITR_Jaune;
266	IPR=Rouge	IPR=Rouge	1	1	0,769	0,083	0,040	1,00039	1,003		1ACID_vert;
275	IPR=Rouge	IBMR=Rouge	4	4	0,661	0,333	0,880		0,194		4PEST_Rouge;
273	IPR=Rouge	IBMR=Rouge	3	3	0,651	0,250	0,960		0,156		2PEST_Rouge;1MPOR_Orange;
274	IPR=Rouge	IBMR=Rouge	3	3	0,722	0,250	0,680		0,123		3PEST_Rouge;
270	IPR=Rouge	IBMR=Rouge	2	2	0,659	0,167	0,960		0,105		2MPOR_Orange;

ANNEXE 5 a: Caractéristiques des cinq motifs sélectionnés par contexte, sur la base de leur combinaison P la plus élevée, extraits pour la longueur de séquences 3 mois, avec les grilles DCE

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E		Items du motif
7	I2M2=Bleu	IPR=Bleu	2	2	0,526	0,333	0,720			0,126	2BILO2_Bleu;
5	I2M2=Bleu	IPR=Bleu	2	2	0,501	0,333	0,600			0,100	1BILO2_Bleu;1SDP_Rouge;
3	I2M2=Bleu	IBD=Bleu	1	1	0,828	0,167	0,160			0,022	1BILO2_Bleu;
2	I2M2=Bleu	IPR=Bleu	1	1	0,482	0,167	0,120			0,010	1POSPE_Bleu;
6	I2M2=Bleu	IBGN=Rouge	2	2	0,549	0,333	0,040			0,007	2SDP_Rouge;
37	I2M2=Vert	IPR=Bleu	2	2	0,507	0,333	0,600			0,101	1BILO2_Bleu;1SDP_Rouge;
39	I2M2=Vert	IBD=Vert	2	1	0,522	0,333	0,200			0,035	1POSPE_Bleu;1SDP_Rouge;
34	I2M2=Vert	IBD=Bleu	1	1	0,754	0,167	0,160			0,020	1BILO2_Bleu;
36	I2M2=Vert	IPR=Bleu	1	1	0,530	0,167	0,120			0,011	1POSPE_Bleu;
38	I2M2=Vert	IBGN=Rouge	2	2	0,566	0,333	0,040			0,008	2SDP_Rouge;
14	I2M2=Jaune	IBD=Vert	2	1	0,548	0,333	0,200			0,037	1POSPE_Bleu;1SDP_Rouge;
12	I2M2=Jaune	IBD=Bleu	1	1	0,630	0,167	0,160			0,017	1BILO2_Bleu;
10	I2M2=Jaune	IPR=Bleu	1	1	0,562	0,167	0,120			0,011	1POSPE_Bleu;
13	I2M2=Jaune	IBGN=Rouge	2	2	0,586	0,333	0,040			0,008	2SDP_Rouge;
9	I2M2=Jaune	IBMR=Orange	1	1	0,575	0,167	0,080			0,008	1BILO2_Vert;
24	I2M2=Orange	IBD=Vert	4	3	0,504	0,667	0,720			0,242	1POSPE_Bleu;3SDP_Rouge;
25	I2M2=Orange	IBD=Vert	3	2	0,483	0,500	0,800			0,193	1POSPE_Bleu;2SDP_Rouge;
23	I2M2=Orange	IBD=Vert	2	2	0,484	0,333	0,920			0,149	1POSPE_Bleu;1SDP_Rouge;
21	I2M2=Orange	IBD=Vert	2	2	0,571	0,333	0,680			0,129	1POSPE_Bleu;1SDP_Rouge;
20	I2M2=Orange	IBD=Vert	2	1	0,569	0,333	0,200			0,038	1POSPE_Bleu;1SDP_Rouge;
31	I2M2=Rouge	IBD=Vert	2	1	0,564	0,333	0,200			0,038	1POSPE_Bleu;1SDP_Rouge;
27	I2M2=Rouge	IPR=Bleu	1	1	0,580	0,167	0,120			0,012	1POSPE_Bleu;
30	I2M2=Rouge	IBGN=Rouge	2	2	0,580	0,333	0,040			0,008	2SDP_Rouge;
28	I2M2=Rouge	IBMR=Orange	1	1	0,567	0,167	0,080			0,008	1BILO2_Vert;
29	I2M2=Rouge	IBGN=Rouge	1	1	0,702	0,167	0,040			0,005	1SDP_Rouge;
46	IBD=Bleu	IBD=Bleu	2	1	0,555	0,333	0,960		1,458	1,636	1BILO2_Bleu;1NUTRI_Bleu;
44	IBD=Bleu	IBD=Bleu	2	2	0,593	0,333	0,960		1,432	1,622	1BILO2_Bleu;1NUTRI_Bleu;
41	IBD=Bleu	IBD=Bleu	1	1	0,655	0,167	0,960		1,432	1,537	1NUTRI_Bleu;
43	IBD=Bleu	IBD=Bleu	1	1	0,864	0,167	0,160		1,044	1,067	1BILO2_Bleu;
45	IBD=Bleu	IPR=Bleu	2	2	0,546	0,333	0,720			0,131	2BILO2_Bleu;
93	IBD=Vert	IBD=Vert	6	4	0,484	1,000	0,960		1,057	1,522	2POSPE_Bleu;4SDP_Rouge;
88	IBD=Vert	IBD=Vert	5	4	0,519	0,833	0,960		1,074	1,489	2POSPE_Bleu;3SDP_Rouge;
91	IBD=Vert	IBD=Vert	5	4	0,484	0,833	0,960		1,058	1,446	2POSPE_Bleu;3SDP_Rouge;
90	IBD=Vert	IBD=Vert	4	3	0,544	0,667	0,720		1,080	1,342	1POSPE_Bleu;3SDP_Rouge;
85	IBD=Vert	IBD=Vert	3	3	0,505	0,500	0,920		1,095	1,328	1BILO2_Bleu;2SDP_Rouge;
60	IBD=Jaune	IBD=Vert	4	3	0,483	0,667	0,920			0,296	2POSPE_Bleu;2SDP_Rouge;
58	IBD=Jaune	IBD=Vert	3	3	0,502	0,500	0,960			0,241	1POSPE_Bleu;2SDP_Rouge;
61	IBD=Jaune	IBD=Vert	4	3	0,502	0,667	0,720			0,241	1POSPE_Bleu;3SDP_Rouge;
62	IBD=Jaune	IBD=Vert	3	3	0,483	0,500	0,960			0,232	1POSPE_Bleu;2SDP_Rouge;
63	IBD=Jaune	IBD=Vert	3	2	0,482	0,500	0,800			0,193	1POSPE_Bleu;2SDP_Rouge;
65	IBD=Orange	IBD=Bleu	1	1	0,563	0,167	0,160			0,015	1BILO2_Bleu;
66	IBD=Orange	IBMR=Orange	1	1	0,582	0,167	0,080			0,008	1BILO2_Vert;
68	IBD=Orange	IBGN=Rouge	2	2	0,568	0,333	0,040			0,008	2SDP_Rouge;
67	IBD=Orange	IBGN=Rouge	1	1	0,657	0,167	0,040			0,004	1SDP_Rouge;
72	IBD=Rouge	IBD=Bleu	1	1	0,705	0,167	0,160			0,019	1BILO2_Bleu;
71	IBD=Rouge	IBMR=Orange	1	1	0,603	0,167	0,080			0,008	1BILO2_Vert;
73	IBD=Rouge	IBGN=Rouge	2	2	0,500	0,333	0,040			0,007	2SDP_Rouge;
70	IBD=Rouge	IBGN=Rouge	1	1	0,564	0,167	0,040			0,004	1SDP_Rouge;
100	IBGN=Bleu	IPR=Bleu	2	2	0,485	0,333	0,720			0,116	2BILO2_Bleu;
103	IBGN=Bleu	IBD=Vert	2	2	0,499	0,333	0,680			0,113	1POSPE_Bleu;1SDP_Rouge;
101	IBGN=Bleu	IPR=Bleu	2	2	0,484	0,333	0,600			0,097	1BILO2_Bleu;1SDP_Rouge;
99	IBGN=Bleu	IBD=Vert	2	1	0,498	0,333	0,200			0,033	1POSPE_Bleu;1SDP_Rouge;
95	IBGN=Bleu	IBD=Bleu	1	1	0,778	0,167	0,160			0,021	1BILO2_Bleu;
130	IBGN=Vert	IPR=Bleu	2	2	0,481	0,333	0,600			0,096	1BILO2_Bleu;1SDP_Rouge;
132	IBGN=Vert	IBD=Vert	2	1	0,531	0,333	0,200			0,035	1POSPE_Bleu;1SDP_Rouge;
128	IBGN=Vert	IBD=Bleu	1	1	0,704	0,167	0,160			0,019	1BILO2_Bleu;
129	IBGN=Vert	IPR=Bleu	1	1	0,543	0,167	0,120			0,011	1POSPE_Bleu;
131	IBGN=Vert	IBGN=Rouge	2	2	0,585	0,333	0,040			0,008	2SDP_Rouge;
110	IBGN=Jaune	IBD=Vert	2	2	0,554	0,333	0,680			0,126	1POSPE_Bleu;1SDP_Rouge;
111	IBGN=Jaune	IBD=Vert	2	1	0,553	0,333	0,200			0,037	1POSPE_Bleu;1SDP_Rouge;
107	IBGN=Jaune	IBD=Bleu	1	1	0,601	0,167	0,160			0,016	1BILO2_Bleu;
106	IBGN=Jaune	IPR=Bleu	1	1	0,573	0,167	0,120			0,011	1POSPE_Bleu;
109	IBGN=Jaune	IBGN=Rouge	2	2	0,588	0,333	0,040			0,008	2SDP_Rouge;

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E		Items du motif
118	IBGN=Orange	IBD=Vert	2	1	0,529	0,333	0,200			0,035	1POSPE_Bleu;1SDP_Rouge;
114	IBGN=Orange	IBD=Bleu	1	1	0,490	0,167	0,160			0,013	1BILO2_Bleu;
115	IBGN=Orange	IPR=Bleu	1	1	0,554	0,167	0,120			0,011	1POSPE_Bleu;
117	IBGN=Orange	IBGN=Rouge	2	2	0,608	0,333	0,040			0,008	2SDP_Rouge;
113	IBGN=Orange	IBMR=Orange	1	1	0,583	0,167	0,080			0,008	1BILO2_Vert;
124	IBGN=Rouge	IBGN=Rouge	3	3	0,593	0,500	0,840		1,202	1,451	3SDP_Rouge;
121	IBGN=Rouge	IBGN=Rouge	1	1	0,815	0,167	0,040		1,139	1,144	1SDP_Rouge;
123	IBGN=Rouge	IBGN=Rouge	2	2	0,704	0,333	0,040		1,099	1,108	2SDP_Rouge;
122	IBGN=Rouge	IBD=Vert	2	1	0,481	0,333	0,200			0,032	1POSPE_Bleu;1SDP_Rouge;
120	IBGN=Rouge	IPR=Bleu	1	1	0,519	0,167	0,120			0,010	1POSPE_Bleu;
139	IBMR=Bleu	IPR=Bleu	2	1	0,488	0,333	0,800			0,130	1BILO2_Bleu;1SDP_Rouge;
138	IBMR=Bleu	IPR=Bleu	2	2	0,510	0,333	0,720			0,123	2BILO2_Bleu;
140	IBMR=Bleu	IBD=Vert	2	2	0,485	0,333	0,680			0,110	1POSPE_Bleu;1SDP_Rouge;
143	IBMR=Bleu	IPR=Bleu	2	2	0,509	0,333	0,600			0,102	1BILO2_Bleu;1SDP_Rouge;
142	IBMR=Bleu	IBD=Vert	2	1	0,484	0,333	0,200			0,032	1POSPE_Bleu;1SDP_Rouge;
169	IBMR=Vert	IPR=Bleu	2	2	0,490	0,333	0,600			0,098	1BILO2_Bleu;1SDP_Rouge;
171	IBMR=Vert	IBD=Vert	2	1	0,504	0,333	0,200			0,034	1POSPE_Bleu;1SDP_Rouge;
167	IBMR=Vert	IBD=Bleu	1	1	0,726	0,167	0,160			0,019	1BILO2_Bleu;
168	IBMR=Vert	IPR=Bleu	1	1	0,520	0,167	0,120			0,010	1POSPE_Bleu;
165	IBMR=Vert	IBMR=Orange	1	1	0,604	0,167	0,080			0,008	1BILO2_Vert;
150	IBMR=Jaune	IBD=Vert	2	2	0,527	0,333	0,680			0,119	1POSPE_Bleu;1SDP_Rouge;
151	IBMR=Jaune	IBD=Vert	2	1	0,523	0,333	0,200			0,035	1POSPE_Bleu;1SDP_Rouge;
147	IBMR=Jaune	IBD=Bleu	1	1	0,565	0,167	0,160			0,015	1BILO2_Bleu;
148	IBMR=Jaune	IPR=Bleu	1	1	0,553	0,167	0,120			0,011	1POSPE_Bleu;
146	IBMR=Jaune	IBMR=Orange	1	1	0,626	0,167	0,080			0,008	1BILO2_Vert;
156	IBMR=Orange	IBMR=Orange	2	2	0,487	0,333	0,960		1,072	1,227	1BILO2_Vert;1SDP_Rouge;
153	IBMR=Orange	IBMR=Orange	1	1	0,704	0,167	0,080		1,097	1,106	1BILO2_Vert;
159	IBMR=Orange	IBD=Vert	4	3	0,519	0,667	0,720			0,249	1POSPE_Bleu;3SDP_Rouge;
160	IBMR=Orange	IBD=Vert	3	2	0,508	0,500	0,800			0,203	1POSPE_Bleu;2SDP_Rouge;
158	IBMR=Orange	IBD=Vert	2	1	0,582	0,333	0,200			0,039	1POSPE_Bleu;1SDP_Rouge;
162	IBMR=Rouge	IPR=Bleu	1	1	0,526	0,167	0,120			0,011	1POSPE_Bleu;
163	IBMR=Rouge	IBMR=Orange	1	1	0,684	0,167	0,080			0,009	1BILO2_Vert;
188	IPR=Bleu	IPR=Bleu	4	3	0,491	0,667	0,920		1,002	1,303	1BILO2_Bleu;3SDP_Rouge;
177	IPR=Bleu	IPR=Bleu	2	2	0,615	0,333	0,720		1,126	1,274	2BILO2_Bleu;
191	IPR=Bleu	IPR=Bleu	3	2	0,525	0,500	0,920		1,031	1,272	1POSPE_Bleu;2SDP_Rouge;
180	IPR=Bleu	IPR=Bleu	2	2	0,498	0,333	0,960		1,108	1,267	1BILO2_Bleu;1SDP_Rouge;
186	IPR=Bleu	IPR=Bleu	3	2	0,484	0,500	0,960		1,031	1,263	1BILO2_Bleu;1POSPE_Bleu;1SDP_Rouge;
228	IPR=Vert	IBD=Vert	4	3	0,480	0,667	0,720			0,231	1POSPE_Bleu;3SDP_Rouge;
225	IPR=Vert	IPR=Bleu	2	1	0,497	0,333	0,800			0,133	1BILO2_Bleu;1SDP_Rouge;
226	IPR=Vert	IBD=Vert	2	2	0,559	0,333	0,680			0,127	1POSPE_Bleu;1SDP_Rouge;
224	IPR=Vert	IPR=Bleu	2	2	0,517	0,333	0,720			0,124	2BILO2_Bleu;
222	IPR=Vert	IPR=Bleu	2	2	0,540	0,333	0,600			0,108	1BILO2_Bleu;1SDP_Rouge;
199	IPR=Jaune	IBD=Vert	2	2	0,537	0,333	0,680			0,122	1POSPE_Bleu;1SDP_Rouge;
200	IPR=Jaune	IBD=Vert	2	1	0,536	0,333	0,200			0,036	1POSPE_Bleu;1SDP_Rouge;
195	IPR=Jaune	IBD=Bleu	1	1	0,683	0,167	0,160			0,018	1BILO2_Bleu;
197	IPR=Jaune	IPR=Bleu	1	1	0,549	0,167	0,120			0,011	1POSPE_Bleu;
194	IPR=Jaune	IBMR=Orange	1	1	0,612	0,167	0,080			0,008	1BILO2_Vert;
209	IPR=Orange	IBD=Vert	4	3	0,496	0,667	0,720			0,238	1POSPE_Bleu;3SDP_Rouge;
208	IPR=Orange	IBGN=Rouge	3	3	0,499	0,500	0,840			0,210	3SDP_Rouge;
207	IPR=Orange	IBD=Vert	2	1	0,566	0,333	0,200			0,038	1POSPE_Bleu;1SDP_Rouge;
205	IPR=Orange	IBD=Bleu	1	1	0,626	0,167	0,160			0,017	1BILO2_Bleu;
202	IPR=Orange	IPR=Bleu	1	1	0,581	0,167	0,120			0,012	1POSPE_Bleu;
215	IPR=Rouge	IBD=Vert	2	1	0,543	0,333	0,200			0,036	1POSPE_Bleu;1SDP_Rouge;
212	IPR=Rouge	IBD=Bleu	1	1	0,579	0,167	0,160			0,015	1BILO2_Bleu;
213	IPR=Rouge	IPR=Bleu	1	1	0,573	0,167	0,120			0,011	1POSPE_Bleu;
216	IPR=Rouge	IBGN=Rouge	2	2	0,599	0,333	0,040			0,008	2SDP_Rouge;
211	IPR=Rouge	IBMR=Orange	1	1	0,593	0,167	0,080			0,008	1BILO2_Vert;

ANNEXE 5 b: Caractéristiques des cinq motifs sélectionnés par contexte, sur la base de leur combinaison P la plus élevée, extraits pour la longueur de séquences 6 mois, avec les grilles DCE

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E		Items du motif
5	I2M2=Bleu	IPR=Bleu	3	3	0,645	0,500	0,760		0,245		3BILO2_Bleu;
4	I2M2=Bleu	IBD=Bleu	2	2	0,825	0,333	0,400		0,110		2BILO2_Bleu;
3	I2M2=Bleu	IBD=Bleu	1	1	0,928	0,167	0,120		0,019		1BILO2_Bleu;
1	I2M2=Bleu	IBGN=Rouge	1	1	0,615	0,167	0,120		0,012		1SPD_Rouge;
2	I2M2=Bleu	IBMR=Rouge	1	1	0,656	0,167	0,040		0,004		1BILO2_Vert;
33	I2M2=Vert	IBD=Bleu	2	2	0,765	0,333	0,400		0,102		2BILO2_Bleu;
32	I2M2=Vert	IBGN=Rouge	2	2	0,641	0,333	0,160		0,034		2SPD_Rouge;
31	I2M2=Vert	IBD=Bleu	1	1	0,874	0,167	0,120		0,017		1BILO2_Bleu;
30	I2M2=Vert	IBGN=Rouge	1	1	0,666	0,167	0,120		0,013		1SPD_Rouge;
29	I2M2=Vert	IBMR=Rouge	1	1	0,658	0,167	0,040		0,004		1BILO2_Vert;
12	I2M2=Jaune	IBGN=Rouge	3	3	0,623	0,500	0,360		0,112		3SPD_Rouge;
11	I2M2=Jaune	IBD=Bleu	2	2	0,640	0,333	0,400		0,085		2BILO2_Bleu;
10	I2M2=Jaune	IBGN=Rouge	2	2	0,670	0,333	0,160		0,036		2SPD_Rouge;
8	I2M2=Jaune	IBD=Bleu	1	1	0,787	0,167	0,120		0,016		1BILO2_Bleu;
9	I2M2=Jaune	IBGN=Rouge	1	1	0,690	0,167	0,120		0,014		1SPD_Rouge;
18	I2M2=Orange	IPR=Bleu	2	2	0,610	0,333	0,920		0,187		1POSPE_Bleu;1SPD_Rouge;
19	I2M2=Orange	IPR=Bleu	2	1	0,609	0,333	0,800		0,162		1POSPE_Bleu;1SPD_Rouge;
21	I2M2=Orange	IBGN=Rouge	3	3	0,657	0,500	0,360		0,118		3SPD_Rouge;
15	I2M2=Orange	IPR=Bleu	1	1	0,631	0,167	0,600		0,063		1POSPE_Bleu;
20	I2M2=Orange	IBGN=Rouge	2	2	0,687	0,333	0,160		0,037		2SPD_Rouge;
27	I2M2=Rouge	IBGN=Rouge	3	3	0,633	0,500	0,360		0,114		3SPD_Rouge;
23	I2M2=Rouge	IPR=Bleu	1	1	0,607	0,167	0,600		0,061		1POSPE_Bleu;
26	I2M2=Rouge	IBGN=Rouge	2	2	0,692	0,333	0,160		0,037		2SPD_Rouge;
25	I2M2=Rouge	IBGN=Rouge	1	1	0,705	0,167	0,120		0,014		1SPD_Rouge;
24	I2M2=Rouge	IBMR=Rouge	1	1	0,779	0,167	0,040		0,005		1BILO2_Vert;
54	IBD=Bleu	IBD=Bleu	5	4	0,601	0,833	0,960	1,402	1,883		3BILO2_Bleu;2NUTRI_Bleu;
52	IBD=Bleu	IBD=Bleu	4	3	0,601	0,667	0,960	1,415	1,800		1BILO2_Bleu;3NUTRI_Bleu;
46	IBD=Bleu	IBD=Bleu	4	3	0,624	0,667	0,960	1,399	1,799		2BILO2_Bleu;2NUTRI_Bleu;
50	IBD=Bleu	IBD=Bleu	4	3	0,622	0,667	0,960	1,344	1,742		3BILO2_Bleu;1NUTRI_Bleu;
51	IBD=Bleu	IBD=Bleu	4	3	0,619	0,667	0,960	1,329	1,725		2BILO2_Bleu;2NUTRI_Bleu;
89	IBD=Vert	IBD=Vert	5	4	0,607	0,833	0,960	1,039	1,524		2POSPE_Bleu;3SPD_Rouge;
94	IBD=Vert	IBD=Vert	4	3	0,604	0,667	0,960	1,111	1,498		1BILO2_Bleu;3SPD_Rouge;
91	IBD=Vert	IBD=Vert	4	4	0,604	0,667	0,960	1,101	1,487		2BILO2_Bleu;2SPD_Rouge;
96	IBD=Vert	IBD=Vert	4	4	0,603	0,667	0,960	1,101	1,487		2BILO2_Bleu;2SPD_Rouge;
93	IBD=Vert	IBD=Vert	3	3	0,625	0,500	0,960	1,111	1,411		1BILO2_Bleu;2SPD_Rouge;
61	IBD=Jaune	IBGN=Rouge	3	3	0,643	0,500	0,360		0,116		3SPD_Rouge;
56	IBD=Jaune	IPR=Bleu	1	1	0,610	0,167	0,600		0,061		1POSPE_Bleu;
60	IBD=Jaune	IBGN=Rouge	2	2	0,683	0,333	0,160		0,036		2SPD_Rouge;
58	IBD=Jaune	IBD=Bleu	1	1	0,731	0,167	0,120		0,015		1BILO2_Bleu;
59	IBD=Jaune	IBGN=Rouge	1	1	0,695	0,167	0,120		0,014		1SPD_Rouge;
67	IBD=Orange	IBGN=Rouge	3	3	0,603	0,500	0,360		0,109		3SPD_Rouge;
66	IBD=Orange	IBGN=Rouge	2	2	0,661	0,333	0,160		0,035		2SPD_Rouge;
65	IBD=Orange	IBD=Bleu	1	1	0,712	0,167	0,120		0,014		1BILO2_Bleu;
64	IBD=Orange	IBGN=Rouge	1	1	0,685	0,167	0,120		0,014		1SPD_Rouge;
63	IBD=Orange	IBMR=Rouge	1	1	0,736	0,167	0,040		0,005		1BILO2_Vert;
71	IBD=Rouge	IBMR=Bleu	2	2	0,603	0,333	0,800		0,161		1BILO2_Bleu;1BILO2_Vert;
72	IBD=Rouge	IBD=Bleu	2	2	0,705	0,333	0,400		0,094		2BILO2_Bleu;
70	IBD=Rouge	IBD=Bleu	1	1	0,885	0,167	0,120		0,018		1BILO2_Bleu;
69	IBD=Rouge	IBMR=Rouge	1	1	0,705	0,167	0,040		0,005		1BILO2_Vert;
102	IBGN=Bleu	IBD=Bleu	2	2	0,776	0,333	0,400		0,103		2BILO2_Bleu;
103	IBGN=Bleu	IBGN=Rouge	2	2	0,612	0,333	0,160		0,033		2SPD_Rouge;
101	IBGN=Bleu	IBD=Bleu	1	1	0,890	0,167	0,120		0,018		1BILO2_Bleu;
100	IBGN=Bleu	IBGN=Rouge	1	1	0,632	0,167	0,120		0,013		1SPD_Rouge;
99	IBGN=Bleu	IBMR=Rouge	1	1	0,672	0,167	0,040		0,004		1BILO2_Vert;
130	IBGN=Vert	IBD=Bleu	2	2	0,719	0,333	0,400		0,096		2BILO2_Bleu;
131	IBGN=Vert	IBGN=Rouge	2	2	0,653	0,333	0,160		0,035		2SPD_Rouge;
129	IBGN=Vert	IBD=Bleu	1	1	0,840	0,167	0,120		0,017		1BILO2_Bleu;
128	IBGN=Vert	IBGN=Rouge	1	1	0,674	0,167	0,120		0,013		1SPD_Rouge;
127	IBGN=Vert	IBMR=Rouge	1	1	0,688	0,167	0,040		0,005		1BILO2_Vert;
111	IBGN=Jaune	IBGN=Rouge	3	3	0,640	0,500	0,360		0,115		3SPD_Rouge;
110	IBGN=Jaune	IBD=Bleu	2	2	0,607	0,333	0,400		0,081		2BILO2_Bleu;
105	IBGN=Jaune	IPR=Bleu	1	1	0,609	0,167	0,600		0,061		1POSPE_Bleu;
109	IBGN=Jaune	IBGN=Rouge	2	2	0,683	0,333	0,160		0,036		2SPD_Rouge;

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E		Items du motif
108	IBGN=Jaune	IBD=Bleu	1	1	0,737	0,167	0,120			0,015	1BILO2_Bleu;
117	IBGN=Orange	IBGN=Rouge	3	3	0,681	0,500	0,360			0,123	3SPD_Rouge;
116	IBGN=Orange	IBGN=Rouge	2	2	0,710	0,333	0,160			0,038	2SPD_Rouge;
114	IBGN=Orange	IBGN=Rouge	1	1	0,725	0,167	0,120			0,014	1SPD_Rouge;
115	IBGN=Orange	IBD=Bleu	1	1	0,643	0,167	0,120			0,013	1BILO2_Bleu;
113	IBGN=Orange	IBMR=Rouge	1	1	0,758	0,167	0,040			0,005	1BILO2_Vert;
125	IBGN=Rouge	IBGN=Rouge	6	6	0,630	1,000	0,960		1,421	2,025	6SPD_Rouge;
124	IBGN=Rouge	IBGN=Rouge	5	5	0,667	0,833	0,960		1,301	1,834	5SPD_Rouge;
123	IBGN=Rouge	IBGN=Rouge	4	4	0,704	0,667	0,960		1,219	1,669	4SPD_Rouge;
119	IBGN=Rouge	IBGN=Rouge	1	1	0,667	0,167	0,960		1,406	1,513	1BILO2_Jaune;
121	IBGN=Rouge	IBGN=Rouge	2	2	0,852	0,333	0,160		1,200	1,245	2SPD_Rouge;
137	IBMR=Bleu	IBMR=Bleu	2	2	0,665	0,333	0,800		1,028	1,205	1BILO2_Bleu;1BILO2_Vert;
139	IBMR=Bleu	IPR=Bleu	3	3	0,639	0,500	0,760			0,243	3BILO2_Bleu;
138	IBMR=Bleu	IBD=Bleu	2	2	0,809	0,333	0,400			0,108	2BILO2_Bleu;
136	IBMR=Bleu	IBGN=Rouge	2	2	0,614	0,333	0,160			0,033	2SPD_Rouge;
135	IBMR=Bleu	IBD=Bleu	1	1	0,921	0,167	0,120			0,018	1BILO2_Bleu;
162	IBMR=Vert	IBMR=Bleu	2	2	0,629	0,333	0,800			0,168	1BILO2_Bleu;1BILO2_Vert;
163	IBMR=Vert	IBD=Bleu	2	2	0,747	0,333	0,400			0,100	2BILO2_Bleu;
161	IBMR=Vert	IBGN=Rouge	2	2	0,640	0,333	0,160			0,034	2SPD_Rouge;
160	IBMR=Vert	IBD=Bleu	1	1	0,875	0,167	0,120			0,017	1BILO2_Bleu;
158	IBMR=Vert	IBGN=Rouge	1	1	0,648	0,167	0,120			0,013	1SPD_Rouge;
145	IBMR=Jaune	IBGN=Rouge	3	3	0,611	0,500	0,360			0,110	3SPD_Rouge;
144	IBMR=Jaune	IBGN=Rouge	2	2	0,636	0,333	0,160			0,034	2SPD_Rouge;
143	IBMR=Jaune	IBD=Bleu	1	1	0,738	0,167	0,120			0,015	1BILO2_Bleu;
142	IBMR=Jaune	IBGN=Rouge	1	1	0,653	0,167	0,120			0,013	1SPD_Rouge;
141	IBMR=Jaune	IBMR=Rouge	1	1	0,806	0,167	0,040			0,005	1BILO2_Vert;
151	IBMR=Orange	IBMR=Rouge	2	2	0,651	0,333	0,920			0,200	2BILO2_Vert;
153	IBMR=Orange	IBGN=Rouge	3	3	0,619	0,500	0,360			0,111	3SPD_Rouge;
150	IBMR=Orange	IPR=Bleu	1	1	0,608	0,167	0,600			0,061	1POSPE_Bleu;
152	IBMR=Orange	IBGN=Rouge	2	2	0,646	0,333	0,160			0,034	2SPD_Rouge;
147	IBMR=Orange	IBD=Bleu	1	1	0,667	0,167	0,120			0,013	1BILO2_Bleu;
156	IBMR=Rouge	IBMR=Rouge	2	2	0,684	0,333	0,920		1,051	1,261	2BILO2_Vert;
155	IBMR=Rouge	IBMR=Rouge	1	1	0,842	0,167	0,040		1,007	1,013	1BILO2_Vert;
180	IPR=Bleu	IPR=Bleu	3	3	0,700	0,500	0,760		1,010	1,276	3BILO2_Bleu;
174	IPR=Bleu	IPR=Bleu	2	2	0,614	0,333	0,960		1,076	1,272	1BILO2_Bleu;1POSPE_Bleu;
171	IPR=Bleu	IPR=Bleu	2	1	0,623	0,333	0,800		1,005	1,171	1POSPE_Bleu;1SPD_Rouge;
165	IPR=Bleu	IPR=Bleu	1	1	0,659	0,167	0,600		1,045	1,111	1POSPE_Bleu;
178	IPR=Bleu	IBD=Vert	4	3	0,600	0,667	0,920			0,368	1POSPE_Bleu;3SPD_Rouge;
215	IPR=Vert	IPR=Bleu	3	3	0,642	0,500	0,760			0,244	3BILO2_Bleu;
212	IPR=Vert	IBD=Vert	2	2	0,628	0,333	0,840			0,176	1BILO2_Bleu;1SPD_Rouge;
216	IPR=Vert	IBGN=Rouge	3	3	0,605	0,500	0,360			0,109	3SPD_Rouge;
213	IPR=Vert	IBD=Bleu	2	2	0,798	0,333	0,400			0,106	2BILO2_Bleu;
209	IPR=Vert	IPR=Bleu	1	1	0,613	0,167	0,600			0,061	1POSPE_Bleu;
188	IPR=Jaune	IBGN=Rouge	3	3	0,603	0,500	0,360			0,108	3SPD_Rouge;
186	IPR=Jaune	IBD=Bleu	2	2	0,681	0,333	0,400			0,091	2BILO2_Bleu;
187	IPR=Jaune	IBGN=Rouge	2	2	0,646	0,333	0,160			0,034	2SPD_Rouge;
185	IPR=Jaune	IBD=Bleu	1	1	0,821	0,167	0,120			0,016	1BILO2_Bleu;
184	IPR=Jaune	IBGN=Rouge	1	1	0,694	0,167	0,120			0,014	1SPD_Rouge;
195	IPR=Orange	IPR=Bleu	2	2	0,621	0,333	0,920			0,191	1POSPE_Bleu;1SPD_Rouge;
197	IPR=Orange	IPR=Bleu	2	1	0,620	0,333	0,800			0,165	1POSPE_Bleu;1SPD_Rouge;
198	IPR=Orange	IBGN=Rouge	3	3	0,672	0,500	0,360			0,121	3SPD_Rouge;
196	IPR=Orange	IBD=Bleu	2	2	0,601	0,333	0,400			0,080	2BILO2_Bleu;
191	IPR=Orange	IPR=Bleu	1	1	0,634	0,167	0,600			0,063	1POSPE_Bleu;
205	IPR=Rouge	IPR=Bleu	2	1	0,601	0,333	0,800			0,160	1POSPE_Bleu;1SPD_Rouge;
206	IPR=Rouge	IBGN=Rouge	3	3	0,630	0,500	0,360			0,113	3SPD_Rouge;
202	IPR=Rouge	IPR=Bleu	1	1	0,625	0,167	0,600			0,062	1POSPE_Bleu;
204	IPR=Rouge	IBGN=Rouge	2	2	0,707	0,333	0,160			0,038	2SPD_Rouge;
203	IPR=Rouge	IBGN=Rouge	1	1	0,760	0,167	0,120			0,015	1SPD_Rouge;

ANNEXE 5 c: Caractéristiques des cinq motifs sélectionnés par contexte, sur la base de leur combinaison P la plus élevée, extraits pour la longueur de séquences 12 mois, avec les grilles DCE

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E		Items du motif
7	I2M2=Bleu	IBMR=Bleu	3	3	0,806	0,429	0,800		0,276		3BILO2_Bleu;
4	I2M2=Bleu	IBMR=Bleu	2	2	0,771	0,286	0,720		0,159		1BILO2_Bleu;1SPD_Rouge;
5	I2M2=Bleu	IBD=Bleu	2	2	0,902	0,286	0,520		0,134		2BILO2_Bleu;
8	I2M2=Bleu	IBGN=Rouge	3	3	0,764	0,429	0,120		0,039		3SPD_Rouge;
3	I2M2=Bleu	IBD=Bleu	1	1	0,956	0,143	0,280		0,038		1BILO2_Bleu;
41	I2M2=Vert	IBD=Bleu	2	2	0,825	0,286	0,520		0,123		2BILO2_Bleu;
43	I2M2=Vert	IBGN=Rouge	3	3	0,768	0,429	0,120		0,039		3SPD_Rouge;
40	I2M2=Vert	IBD=Bleu	1	1	0,908	0,143	0,280		0,036		1BILO2_Bleu;
42	I2M2=Vert	IBGN=Rouge	2	2	0,799	0,286	0,080		0,018		2SPD_Rouge;
39	I2M2=Vert	IBGN=Rouge	1	1	0,815	0,143	0,080		0,009		1SPD_Rouge;
15	I2M2=Jaune	IBGN=Rouge	4	4	0,761	0,571	0,520		0,226		4SPD_Rouge;
14	I2M2=Jaune	IBGN=Rouge	3	3	0,806	0,429	0,120		0,041		3SPD_Rouge;
12	I2M2=Jaune	IBD=Bleu	1	1	0,815	0,143	0,280		0,033		1BILO2_Bleu;
13	I2M2=Jaune	IBGN=Rouge	2	2	0,834	0,286	0,080		0,019		2SPD_Rouge;
11	I2M2=Jaune	IBGN=Rouge	1	1	0,842	0,143	0,080		0,010		1SPD_Rouge;
23	I2M2=Orange	IBMR=Orange	3	3	0,761	0,429	0,920		0,300		1POSPE_Bleu;2SPD_Rouge;
25	I2M2=Orange	IBGN=Rouge	4	4	0,784	0,571	0,520		0,233		4SPD_Rouge;
21	I2M2=Orange	IBMR=Orange	2	2	0,771	0,286	0,840		0,185		1POSPE_Bleu;1SPD_Rouge;
20	I2M2=Orange	IBMR=Orange	2	1	0,767	0,286	0,840		0,184		1POSPE_Bleu;1SPD_Rouge;
18	I2M2=Orange	IBMR=Orange	1	1	0,772	0,143	0,840		0,093		1POSPE_Bleu;
32	I2M2=Rouge	IBMR=Orange	4	3	0,763	0,571	0,920		0,401		1POSPE_Bleu;3SPD_Rouge;
35	I2M2=Rouge	IBMR=Orange	4	4	0,761	0,571	0,880		0,383		1BILO2_Vert;3SPD_Rouge;
33	I2M2=Rouge	IBMR=Orange	3	3	0,777	0,429	0,840		0,280		1BILO2_Vert;2SPD_Rouge;
36	I2M2=Rouge	IBGN=Rouge	4	4	0,796	0,571	0,520		0,236		4SPD_Rouge;
30	I2M2=Rouge	IBMR=Orange	2	2	0,780	0,286	0,840		0,187		1BILO2_Vert;1SPD_Rouge;
45	IBD=Bleu	IBD=Bleu	1	1	0,768	0,143	0,960	1,275	1,380		1NUTRI_Bleu;
48	IBD=Bleu	IBD=Bleu	2	2	0,916	0,286	0,520	1,016	1,152		2BILO2_Bleu;
47	IBD=Bleu	IBD=Bleu	1	1	0,968	0,143	0,280	1,012	1,051		1BILO2_Bleu;
51	IBD=Bleu	IBMR=Bleu	3	3	0,822	0,429	0,800		0,282		3BILO2_Bleu;
49	IBD=Bleu	IBMR=Bleu	2	2	0,767	0,286	0,720		0,158		1BILO2_Bleu;1SPD_Rouge;
81	IBD=Vert	IBMR=Bleu	3	3	0,766	0,429	0,880		0,289		1BILO2_Bleu;2SPD_Rouge;
83	IBD=Vert	IBMR=Bleu	3	3	0,776	0,429	0,800		0,266		1BILO2_Bleu;2SPD_Rouge;
84	IBD=Vert	IBGN=Rouge	4	4	0,770	0,571	0,520		0,229		4SPD_Rouge;
76	IBD=Vert	IBMR=Bleu	2	2	0,777	0,286	0,880		0,195		1BILO2_Bleu;1BILO2_Vert;
78	IBD=Vert	IBMR=Bleu	2	2	0,768	0,286	0,880		0,193		1BILO2_Bleu;1SPD_Rouge;
57	IBD=Jaune	IBGN=Rouge	3	3	0,795	0,429	0,120		0,041		3SPD_Rouge;
55	IBD=Jaune	IBD=Bleu	1	1	0,772	0,143	0,280		0,031		1BILO2_Bleu;
56	IBD=Jaune	IBGN=Rouge	2	2	0,820	0,286	0,080		0,019		2SPD_Rouge;
54	IBD=Jaune	IBGN=Rouge	1	1	0,828	0,143	0,080		0,009		1SPD_Rouge;
53	IBD=Jaune	IBMR=Orange	1	1	0,870	0,143	0,040		0,005		1BILO2_Vert;
62	IBD=Orange	IBGN=Rouge	3	3	0,783	0,429	0,120		0,040		3SPD_Rouge;
61	IBD=Orange	IBGN=Rouge	2	2	0,811	0,286	0,080		0,019		2SPD_Rouge;
60	IBD=Orange	IBGN=Rouge	1	1	0,817	0,143	0,080		0,009		1SPD_Rouge;
59	IBD=Orange	IBMR=Orange	1	1	0,839	0,143	0,040		0,005		1BILO2_Vert;
69	IBD=Rouge	IBGN=Rouge	4	4	0,768	0,571	0,520		0,228		4SPD_Rouge;
67	IBD=Rouge	IBD=Bleu	2	2	0,780	0,286	0,520		0,116		2BILO2_Bleu;
68	IBD=Rouge	IBGN=Rouge	3	3	0,805	0,429	0,120		0,041		3SPD_Rouge;
65	IBD=Rouge	IBD=Bleu	1	1	0,902	0,143	0,280		0,036		1BILO2_Bleu;
66	IBD=Rouge	IBGN=Rouge	2	2	0,829	0,286	0,080		0,019		2SPD_Rouge;
89	IBGN=Bleu	IBD=Bleu	2	2	0,841	0,286	0,520		0,125		2BILO2_Bleu;
91	IBGN=Bleu	IBGN=Rouge	3	3	0,767	0,429	0,120		0,039		3SPD_Rouge;
88	IBGN=Bleu	IBD=Bleu	1	1	0,919	0,143	0,280		0,037		1BILO2_Bleu;
90	IBGN=Bleu	IBGN=Rouge	2	2	0,794	0,286	0,080		0,018		2SPD_Rouge;
87	IBGN=Bleu	IBGN=Rouge	1	1	0,807	0,143	0,080		0,009		1SPD_Rouge;
118	IBGN=Vert	IBD=Bleu	2	2	0,779	0,286	0,520		0,116		2BILO2_Bleu;
119	IBGN=Vert	IBGN=Rouge	3	3	0,785	0,429	0,120		0,040		3SPD_Rouge;
114	IBGN=Vert	IBD=Bleu	1	1	0,867	0,143	0,280		0,035		1BILO2_Bleu;
117	IBGN=Vert	IBGN=Rouge	2	2	0,816	0,286	0,080		0,019		2SPD_Rouge;
115	IBGN=Vert	IBGN=Rouge	1	1	0,829	0,143	0,080		0,009		1SPD_Rouge;
98	IBGN=Jaune	IBGN=Rouge	4	4	0,783	0,571	0,520		0,233		4SPD_Rouge;
97	IBGN=Jaune	IBGN=Rouge	3	3	0,818	0,429	0,120		0,042		3SPD_Rouge;
94	IBGN=Jaune	IBD=Bleu	1	1	0,786	0,143	0,280		0,031		1BILO2_Bleu;
96	IBGN=Jaune	IBGN=Rouge	2	2	0,847	0,286	0,080		0,019		2SPD_Rouge;

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E		Items du motif
95	IBGN=Jaune	IBGN=Rouge	1	1	0,857	0,143	0,080			0,010	1SPD_Rouge;
104	IBGN=Orange	IBMR=Orange	3	3	0,761	0,429	0,840			0,274	1BILO2_Vert;2SPD_Rouge;
106	IBGN=Orange	IBGN=Rouge	4	4	0,818	0,571	0,520			0,243	4SPD_Rouge;
102	IBGN=Orange	IBMR=Orange	2	2	0,765	0,286	0,840			0,184	1BILO2_Vert;1SPD_Rouge;
105	IBGN=Orange	IBGN=Rouge	3	3	0,850	0,429	0,120			0,044	3SPD_Rouge;
103	IBGN=Orange	IBGN=Rouge	2	2	0,874	0,286	0,080			0,020	2SPD_Rouge;
111	IBGN=Rouge	IBGN=Rouge	4	4	0,933	0,571	0,520		1,141	1,419	4SPD_Rouge;
108	IBGN=Rouge	IBGN=Rouge	1	1	0,767	0,143	0,920		1,262	1,363	1BILO2_Jaune;
110	IBGN=Rouge	IBGN=Rouge	1	1	0,967	0,143	0,080		1,095	1,106	1SPD_Rouge;
109	IBGN=Rouge	IBMR=Orange	1	1	0,767	0,143	0,040			0,004	1BILO2_Vert;
148	IBMR=Bleu	IBMR=Bleu	5	5	0,788	0,714	0,960		1,102	1,643	1BILO2_Bleu;4SPD_Rouge;
150	IBMR=Bleu	IBMR=Bleu	5	5	0,774	0,714	0,960		1,106	1,637	1BILO2_Bleu;4SPD_Rouge;
149	IBMR=Bleu	IBMR=Bleu	5	5	0,781	0,714	0,960		1,100	1,636	1BILO2_Bleu;4SPD_Rouge;
147	IBMR=Bleu	IBMR=Bleu	5	5	0,766	0,714	0,960		1,092	1,617	2BILO2_Bleu;3SPD_Rouge;
146	IBMR=Bleu	IBMR=Bleu	5	5	0,763	0,714	0,960		1,092	1,615	2BILO2_Bleu;3SPD_Rouge;
216	IBMR=Vert	IBMR=Bleu	3	3	0,761	0,429	0,800			0,261	1BILO2_Bleu;2SPD_Rouge;
218	IBMR=Vert	IBGN=Rouge	4	4	0,812	0,571	0,520			0,241	4SPD_Rouge;
213	IBMR=Vert	IBMR=Bleu	2	2	0,783	0,286	0,880			0,197	1BILO2_Bleu;1BILO2_Vert;
214	IBMR=Vert	IBMR=Bleu	2	2	0,765	0,286	0,720			0,157	1BILO2_Bleu;1SPD_Rouge;
212	IBMR=Vert	IBD=Bleu	2	2	0,838	0,286	0,520			0,124	2BILO2_Bleu;
162	IBMR=Jaune	IBMR=Orange	5	5	0,772	0,714	0,920			0,507	1BILO2_Vert;4SPD_Rouge;
160	IBMR=Jaune	IBMR=Orange	4	4	0,781	0,571	0,880			0,393	1BILO2_Vert;3SPD_Rouge;
158	IBMR=Jaune	IBMR=Orange	3	3	0,787	0,429	0,840			0,283	1BILO2_Vert;2SPD_Rouge;
161	IBMR=Jaune	IBGN=Rouge	4	4	0,833	0,571	0,520			0,247	4SPD_Rouge;
155	IBMR=Jaune	IBMR=Rouge	2	2	0,790	0,286	0,880			0,199	2BILO2_Vert;
200	IBMR=Orange	IBMR=Orange	7	5	0,760	1,000	0,960		1,047	1,777	2POSPE_Bleu;5SPD_Rouge;
195	IBMR=Orange	IBMR=Orange	6	5	0,760	0,857	0,960		1,047	1,673	1POSPE_Bleu;5SPD_Rouge;
191	IBMR=Orange	IBMR=Orange	6	4	0,766	0,857	0,960		1,033	1,663	2POSPE_Bleu;4SPD_Rouge;
197	IBMR=Orange	IBMR=Orange	5	5	0,760	0,714	0,960		1,074	1,596	1BILO2_Vert;4SPD_Rouge;
193	IBMR=Orange	IBMR=Orange	5	5	0,776	0,714	0,960		1,063	1,596	1BILO2_Vert;4SPD_Rouge;
234	IPR=Bleu	IBMR=Bleu	4	4	0,765	0,571	0,920			0,402	1BILO2_Bleu;3SPD_Rouge;
233	IPR=Bleu	IBMR=Bleu	3	3	0,773	0,429	0,880			0,291	1BILO2_Bleu;2SPD_Rouge;
230	IPR=Bleu	IBMR=Bleu	3	3	0,794	0,429	0,800			0,272	3BILO2_Bleu;
231	IPR=Bleu	IBMR=Bleu	3	3	0,780	0,429	0,800			0,268	1BILO2_Bleu;2SPD_Rouge;
229	IPR=Bleu	IBMR=Bleu	2	2	0,777	0,286	0,880			0,195	1BILO2_Bleu;1SPD_Rouge;
262	IPR=Vert	IBMR=Bleu	3	3	0,762	0,429	0,800			0,261	1BILO2_Bleu;2SPD_Rouge;
261	IPR=Vert	IBMR=Bleu	3	3	0,760	0,429	0,800			0,261	3BILO2_Bleu;
259	IPR=Vert	IBMR=Bleu	2	2	0,776	0,286	0,720			0,160	1BILO2_Bleu;1SPD_Rouge;
258	IPR=Vert	IBD=Bleu	2	2	0,852	0,286	0,520			0,127	2BILO2_Bleu;
263	IPR=Vert	IBGN=Rouge	3	3	0,800	0,429	0,120			0,041	3SPD_Rouge;
239	IPR=Jaune	IBD=Bleu	2	2	0,764	0,286	0,520			0,113	2BILO2_Bleu;
241	IPR=Jaune	IBGN=Rouge	3	3	0,794	0,429	0,120			0,041	3SPD_Rouge;
237	IPR=Jaune	IBD=Bleu	1	1	0,873	0,143	0,280			0,035	1BILO2_Bleu;
240	IPR=Jaune	IBGN=Rouge	2	2	0,819	0,286	0,080			0,019	2SPD_Rouge;
238	IPR=Jaune	IBGN=Rouge	1	1	0,840	0,143	0,080			0,010	1SPD_Rouge;
247	IPR=Orange	IBGN=Rouge	3	3	0,816	0,429	0,120			0,042	3SPD_Rouge;
244	IPR=Orange	IBD=Bleu	1	1	0,789	0,143	0,280			0,032	1BILO2_Bleu;
246	IPR=Orange	IBGN=Rouge	2	2	0,850	0,286	0,080			0,019	2SPD_Rouge;
245	IPR=Orange	IBGN=Rouge	1	1	0,873	0,143	0,080			0,010	1SPD_Rouge;
243	IPR=Orange	IBMR=Orange	1	1	0,828	0,143	0,040			0,005	1BILO2_Vert;
253	IPR=Rouge	IBGN=Rouge	3	3	0,833	0,429	0,120			0,043	3SPD_Rouge;
250	IPR=Rouge	IBD=Bleu	1	1	0,822	0,143	0,280			0,033	1BILO2_Bleu;
252	IPR=Rouge	IBGN=Rouge	2	2	0,857	0,286	0,080			0,020	2SPD_Rouge;
251	IPR=Rouge	IBGN=Rouge	1	1	0,883	0,143	0,080			0,010	1SPD_Rouge;
249	IPR=Rouge	IBMR=Orange	1	1	0,830	0,143	0,040			0,005	1BILO2_Vert;
7	I2M2=Bleu	IBMR=Bleu	3	3	0,806	0,429	0,800			0,276	3BILO2_Bleu;
4	I2M2=Bleu	IBMR=Bleu	2	2	0,771	0,286	0,720			0,159	1BILO2_Bleu;1SPD_Rouge;
5	I2M2=Bleu	IBD=Bleu	2	2	0,902	0,286	0,520			0,134	2BILO2_Bleu;
8	I2M2=Bleu	IBGN=Rouge	3	3	0,764	0,429	0,120			0,039	3SPD_Rouge;
3	I2M2=Bleu	IBD=Bleu	1	1	0,956	0,143	0,280			0,038	1BILO2_Bleu;

ANNEXE 5 d: Caractéristiques des cinq motifs sélectionnés par contexte, sur la base de leur combinaison P la plus élevée, extraits pour la longueur de séquences 18 mois, avec les grilles DCE

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E		Items du motif
7	I2M2=Bleu	IBD=Bleu	4	4	0,839	0,667	0,760			0,425	4BILO2_Bleu;
6	I2M2=Bleu	IBD=Bleu	3	3	0,898	0,500	0,640			0,287	3BILO2_Bleu;
5	I2M2=Bleu	IBD=Bleu	2	2	0,942	0,333	0,520			0,163	2BILO2_Bleu;
3	I2M2=Bleu	IBD=Bleu	1	1	0,967	0,167	0,280			0,045	1BILO2_Bleu;
4	I2M2=Bleu	IBGN=Rouge	2	2	0,805	0,333	0,120			0,032	2SPD_Rouge;
41	I2M2=Vert	IBD=Bleu	3	3	0,835	0,500	0,640			0,267	3BILO2_Bleu;
40	I2M2=Vert	IBD=Bleu	2	2	0,893	0,333	0,520			0,155	2BILO2_Bleu;
42	I2M2=Vert	IBGN=Rouge	3	3	0,804	0,500	0,200			0,080	3SPD_Rouge;
36	I2M2=Vert	IBD=Bleu	1	1	0,938	0,167	0,280			0,044	1BILO2_Bleu;
39	I2M2=Vert	IBGN=Rouge	2	2	0,817	0,333	0,120			0,033	2SPD_Rouge;
15	I2M2=Jaune	IBGN=Rouge	5	5	0,801	0,833	0,360			0,240	5SPD_Rouge;
14	I2M2=Jaune	IBGN=Rouge	4	4	0,825	0,667	0,200			0,110	4SPD_Rouge;
13	I2M2=Jaune	IBGN=Rouge	3	3	0,836	0,500	0,200			0,084	3SPD_Rouge;
11	I2M2=Jaune	IBD=Bleu	1	1	0,848	0,167	0,280			0,040	1BILO2_Bleu;
12	I2M2=Jaune	IBGN=Rouge	2	2	0,844	0,333	0,120			0,034	2SPD_Rouge;
24	I2M2=Orange	IBGN=Rouge	6	6	0,800	1,000	0,680			0,544	6SPD_Rouge;
23	I2M2=Orange	IBGN=Rouge	5	5	0,834	0,833	0,360			0,250	5SPD_Rouge;
19	I2M2=Orange	IBMR=Rouge	2	2	0,803	0,333	0,760			0,203	2BILO2_Vert;
22	I2M2=Orange	IBGN=Rouge	4	4	0,855	0,667	0,200			0,114	4SPD_Rouge;
21	I2M2=Orange	IBGN=Rouge	3	3	0,860	0,500	0,200			0,086	3SPD_Rouge;
34	I2M2=Rouge	IBGN=Rouge	6	6	0,826	1,000	0,680			0,562	6SPD_Rouge;
31	I2M2=Rouge	IBGN=Rouge	4	4	0,807	0,667	0,880			0,474	1BILO2_Vert;3SPD_Rouge;
29	I2M2=Rouge	IBGN=Rouge	3	3	0,810	0,500	0,920			0,373	1BILO2_Vert;2SPD_Rouge;
33	I2M2=Rouge	IBGN=Rouge	5	5	0,842	0,833	0,360			0,253	5SPD_Rouge;
28	I2M2=Rouge	IBMR=Rouge	2	2	0,818	0,333	0,760			0,207	2BILO2_Vert;
50	IBD=Bleu	IBD=Bleu	4	4	0,850	0,667	0,760		1,012	1,443	4BILO2_Bleu;
49	IBD=Bleu	IBD=Bleu	3	3	0,906	0,500	0,640		1,009	1,299	3BILO2_Bleu;
46	IBD=Bleu	IBD=Bleu	2	2	0,947	0,333	0,520		1,005	1,169	2BILO2_Bleu;
45	IBD=Bleu	IBD=Bleu	1	1	0,974	0,167	0,280		1,007	1,052	1BILO2_Bleu;
47	IBD=Bleu	IBMR=Bleu	2	2	0,803	0,333	0,800			0,214	1BILO2_Bleu;1SPD_Rouge;
78	IBD=Vert	IBD=Vert	2	2	0,816	0,333	0,880		1,085	1,324	1BILO2_Bleu;1BILO2_Vert;
84	IBD=Vert	IBD=Vert	2	2	0,811	0,333	0,880		1,065	1,303	1BILO2_Bleu;1BILO2_Vert;
90	IBD=Vert	IBGN=Rouge	6	6	0,803	1,000	0,680			0,546	6SPD_Rouge;
88	IBD=Vert	IBMR=Bleu	4	4	0,803	0,667	0,920			0,492	1BILO2_Bleu;3SPD_Rouge;
83	IBD=Vert	IBMR=Bleu	3	3	0,810	0,500	0,920			0,373	2BILO2_Bleu;1BILO2_Vert;
55	IBD=Jaune	IBMR=Rouge	2	2	0,828	0,333	0,760			0,210	2BILO2_Vert;
58	IBD=Jaune	IBGN=Rouge	4	4	0,813	0,667	0,200			0,108	4SPD_Rouge;
57	IBD=Jaune	IBGN=Rouge	3	3	0,823	0,500	0,200			0,082	3SPD_Rouge;
54	IBD=Jaune	IBD=Bleu	1	1	0,810	0,167	0,280			0,038	1BILO2_Bleu;
56	IBD=Jaune	IBGN=Rouge	2	2	0,832	0,333	0,120			0,033	2SPD_Rouge;
64	IBD=Orange	IBGN=Rouge	4	4	0,805	0,667	0,200			0,107	4SPD_Rouge;
63	IBD=Orange	IBGN=Rouge	3	3	0,809	0,500	0,200			0,081	3SPD_Rouge;
62	IBD=Orange	IBGN=Rouge	2	2	0,821	0,333	0,120			0,033	2SPD_Rouge;
61	IBD=Orange	IBGN=Rouge	1	1	0,831	0,167	0,200			0,028	1SPD_Rouge;
60	IBD=Orange	IBMR=Rouge	1	1	0,888	0,167	0,040			0,006	1BILO2_Vert;
69	IBD=Rouge	IBD=Bleu	2	2	0,878	0,333	0,520			0,152	2BILO2_Bleu;
71	IBD=Rouge	IBGN=Rouge	4	4	0,805	0,667	0,200			0,107	4SPD_Rouge;
70	IBD=Rouge	IBGN=Rouge	3	3	0,817	0,500	0,200			0,082	3SPD_Rouge;
67	IBD=Rouge	IBD=Bleu	1	1	0,915	0,167	0,280			0,043	1BILO2_Bleu;
68	IBD=Rouge	IBGN=Rouge	2	2	0,829	0,333	0,120			0,033	2SPD_Rouge;
97	IBGN=Bleu	IBD=Bleu	3	3	0,842	0,500	0,640			0,270	3BILO2_Bleu;
96	IBGN=Bleu	IBD=Bleu	2	2	0,902	0,333	0,520			0,156	2BILO2_Bleu;
94	IBGN=Bleu	IBD=Bleu	1	1	0,944	0,167	0,280			0,044	1BILO2_Bleu;
95	IBGN=Bleu	IBGN=Rouge	2	2	0,805	0,333	0,120			0,032	2SPD_Rouge;
93	IBGN=Bleu	IBGN=Rouge	1	1	0,816	0,167	0,200			0,027	1SPD_Rouge;
131	IBGN=Vert	IBD=Bleu	2	2	0,845	0,333	0,520			0,146	2BILO2_Bleu;
134	IBGN=Vert	IBGN=Rouge	4	4	0,811	0,667	0,200			0,108	4SPD_Rouge;
133	IBGN=Vert	IBGN=Rouge	3	3	0,825	0,500	0,200			0,083	3SPD_Rouge;
130	IBGN=Vert	IBD=Bleu	1	1	0,894	0,167	0,280			0,042	1BILO2_Bleu;
132	IBGN=Vert	IBGN=Rouge	2	2	0,837	0,333	0,120			0,033	2SPD_Rouge;
105	IBGN=Jaune	IBGN=Rouge	5	5	0,817	0,833	0,360			0,245	5SPD_Rouge;
104	IBGN=Jaune	IBGN=Rouge	4	4	0,845	0,667	0,200			0,113	4SPD_Rouge;
103	IBGN=Jaune	IBGN=Rouge	3	3	0,856	0,500	0,200			0,086	3SPD_Rouge;

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E		Items du motif
99	IBGN=Jaune	IBD=Bleu	1	1	0,815	0,167	0,280			0,038	1BILO2_Bleu;
102	IBGN=Jaune	IBGN=Rouge	2	2	0,860	0,333	0,120			0,034	2SPD_Rouge;
113	IBGN=Orange	IBGN=Rouge	6	6	0,815	1,000	0,680			0,554	6SPD_Rouge;
112	IBGN=Orange	IBGN=Rouge	5	5	0,839	0,833	0,360			0,252	5SPD_Rouge;
111	IBGN=Orange	IBGN=Rouge	4	4	0,867	0,667	0,200			0,116	4SPD_Rouge;
110	IBGN=Orange	IBGN=Rouge	3	3	0,875	0,500	0,200			0,088	3SPD_Rouge;
109	IBGN=Orange	IBGN=Rouge	2	2	0,883	0,333	0,120			0,035	2SPD_Rouge;
121	IBGN=Rouge	IBGN=Rouge	5	5	0,833	0,833	0,880		1,049	1,660	1BILO2_Vert;4SPD_Rouge;
125	IBGN=Rouge	IBGN=Rouge	6	6	0,867	1,000	0,680		1,049	1,638	6SPD_Rouge;
122	IBGN=Rouge	IBGN=Rouge	4	4	0,833	0,667	0,960		1,075	1,608	1BILO2_Vert;3SPD_Rouge;
118	IBGN=Rouge	IBGN=Rouge	4	4	0,867	0,667	0,880		1,073	1,582	1BILO2_Vert;3SPD_Rouge;
123	IBGN=Rouge	IBGN=Rouge	5	5	0,900	0,833	0,360		1,069	1,339	5SPD_Rouge;
168	IBMR=Bleu	IBMR=Bleu	6	6	0,802	1,000	0,960		1,031	1,801	1BILO2_Bleu;5SPD_Rouge;
165	IBMR=Bleu	IBMR=Bleu	5	5	0,812	0,833	0,960		1,024	1,673	1BILO2_Bleu;4SPD_Rouge;
166	IBMR=Bleu	IBMR=Bleu	5	5	0,801	0,833	0,960		1,027	1,668	1BILO2_Bleu;4SPD_Rouge;
163	IBMR=Bleu	IBMR=Bleu	5	5	0,804	0,833	0,960		1,023	1,666	1BILO2_Bleu;4SPD_Rouge;
157	IBMR=Bleu	IBMR=Bleu	4	4	0,801	0,667	0,960		1,025	1,538	1BILO2_Bleu;3SPD_Rouge;
215	IBMR=Vert	IBD=Bleu	3	3	0,825	0,500	0,640			0,264	3BILO2_Bleu;
219	IBMR=Vert	IBGN=Rouge	5	5	0,820	0,833	0,360			0,246	5SPD_Rouge;
213	IBMR=Vert	IBD=Vert	2	2	0,806	0,333	0,880			0,236	1BILO2_Bleu;1BILO2_Vert;
217	IBMR=Vert	IBD=Vert	2	2	0,806	0,333	0,880			0,236	1BILO2_Bleu;1BILO2_Vert;
214	IBMR=Vert	IBMR=Bleu	2	2	0,838	0,333	0,840			0,235	1BILO2_Bleu;1BILO2_Vert;
180	IBMR=Jaune	IBGN=Rouge	5	5	0,801	0,833	0,880			0,588	1BILO2_Vert;4SPD_Rouge;
182	IBMR=Jaune	IBGN=Rouge	6	6	0,801	1,000	0,680			0,545	6SPD_Rouge;
178	IBMR=Jaune	IBGN=Rouge	4	4	0,805	0,667	0,880			0,472	1BILO2_Vert;3SPD_Rouge;
176	IBMR=Jaune	IBGN=Rouge	3	3	0,809	0,500	0,920			0,372	1BILO2_Vert;2SPD_Rouge;
174	IBMR=Jaune	IBGN=Rouge	2	2	0,811	0,333	0,960			0,259	1BILO2_Vert;1SPD_Rouge;
199	IBMR=Orange	IBMR=Orange	6	6	0,802	1,000	0,960		1,075	1,845	1BILO2_Vert;5SPD_Rouge;
198	IBMR=Orange	IBMR=Orange	6	6	0,802	1,000	0,960		1,060	1,830	2BILO2_Vert;4SPD_Rouge;
197	IBMR=Orange	IBMR=Orange	6	6	0,807	1,000	0,960		1,052	1,827	1BILO2_Vert;5SPD_Rouge;
201	IBMR=Orange	IBMR=Orange	6	6	0,802	1,000	0,960		1,041	1,811	1BILO2_Vert;5SPD_Rouge;
194	IBMR=Orange	IBMR=Orange	6	6	0,818	1,000	0,920		1,022	1,774	1BILO2_Vert;5SPD_Rouge;
205	IBMR=Rouge	IBMR=Rouge	2	2	0,905	0,333	0,760		1,040	1,269	2BILO2_Vert;
203	IBMR=Rouge	IBMR=Rouge	1	1	0,810	0,167	0,920		1,012	1,136	1BILO2_Jaune;
204	IBMR=Rouge	IBMR=Rouge	1	1	0,952	0,167	0,040		1,016	1,022	1BILO2_Vert;
206	IBMR=Rouge	IBMR=Orange	3	3	0,810	0,500	0,920			0,372	3BILO2_Vert;
231	IPR=Bleu	IBD=Bleu	4	4	0,839	0,667	0,760			0,425	4BILO2_Bleu;
230	IPR=Bleu	IBD=Bleu	3	3	0,875	0,500	0,640			0,280	3BILO2_Bleu;
233	IPR=Bleu	IBGN=Rouge	5	5	0,826	0,833	0,360			0,248	5SPD_Rouge;
224	IPR=Bleu	IBMR=Bleu	2	2	0,815	0,333	0,840			0,228	1BILO2_Bleu;1BILO2_Vert;
226	IPR=Bleu	IBMR=Bleu	2	2	0,803	0,333	0,840			0,225	1BILO2_Bleu;1SPD_Rouge;
269	IPR=Vert	IBD=Bleu	4	4	0,802	0,667	0,760			0,407	4BILO2_Bleu;
268	IPR=Vert	IBD=Bleu	3	3	0,863	0,500	0,640			0,276	3BILO2_Bleu;
271	IPR=Vert	IBGN=Rouge	5	5	0,811	0,833	0,360			0,243	5SPD_Rouge;
266	IPR=Vert	IBMR=Bleu	2	2	0,801	0,333	0,840			0,224	1BILO2_Bleu;1SPD_Rouge;
264	IPR=Vert	IBMR=Bleu	2	2	0,809	0,333	0,800			0,216	1BILO2_Bleu;1SPD_Rouge;
242	IPR=Jaune	IBGN=Rouge	5	5	0,815	0,833	0,360			0,245	5SPD_Rouge;
238	IPR=Jaune	IBD=Bleu	2	2	0,845	0,333	0,520			0,146	2BILO2_Bleu;
241	IPR=Jaune	IBGN=Rouge	4	4	0,823	0,667	0,200			0,110	4SPD_Rouge;
240	IPR=Jaune	IBGN=Rouge	3	3	0,834	0,500	0,200			0,083	3SPD_Rouge;
236	IPR=Jaune	IBD=Bleu	1	1	0,902	0,167	0,280			0,042	1BILO2_Bleu;
250	IPR=Orange	IBGN=Rouge	5	5	0,842	0,833	0,360			0,253	5SPD_Rouge;
249	IPR=Orange	IBGN=Rouge	4	4	0,857	0,667	0,200			0,114	4SPD_Rouge;
248	IPR=Orange	IBGN=Rouge	3	3	0,868	0,500	0,200			0,087	3SPD_Rouge;
244	IPR=Orange	IBD=Bleu	1	1	0,835	0,167	0,280			0,039	1BILO2_Bleu;
247	IPR=Orange	IBGN=Rouge	2	2	0,876	0,333	0,120			0,035	2SPD_Rouge;
258	IPR=Rouge	IBGN=Rouge	5	5	0,828	0,833	0,360			0,248	5SPD_Rouge;
257	IPR=Rouge	IBGN=Rouge	4	4	0,857	0,667	0,200			0,114	4SPD_Rouge;
256	IPR=Rouge	IBGN=Rouge	3	3	0,881	0,500	0,200			0,088	3SPD_Rouge;
252	IPR=Rouge	IBD=Bleu	1	1	0,849	0,167	0,280			0,040	1BILO2_Bleu;
255	IPR=Rouge	IBGN=Rouge	2	2	0,892	0,333	0,120			0,036	2SPD_Rouge;

ANNEXE 5 e: Caractéristiques des cinq motifs sélectionnés par contexte, sur la base de leur combinaison P la plus élevée, pour la longueur de séquences 24 mois, avec les grilles DCE

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E		Items du motif
11	I2M2=Bleu	IBMR=Bleu	4	4	0,880	0,444	0,800			0,313	1BILO2_Bleu;3SPD_Rouge;
7	I2M2=Bleu	IBD=Bleu	3	3	0,916	0,333	0,800			0,244	3BILO2_Bleu;
8	I2M2=Bleu	IBMR=Bleu	3	3	0,887	0,333	0,800			0,237	1BILO2_Bleu;2SPD_Rouge;
9	I2M2=Bleu	IBMR=Bleu	3	3	0,884	0,333	0,800			0,236	1BILO2_Bleu;2SPD_Rouge;
4	I2M2=Bleu	IBMR=Bleu	2	2	0,894	0,222	0,760			0,151	1BILO2_Bleu;1SPD_Rouge;
39	I2M2=Vert	IBD=Bleu	2	2	0,906	0,222	0,640			0,129	2BILO2_Bleu;
41	I2M2=Vert	IBMR=Rouge	3	3	0,906	0,333	0,200			0,060	3SPD_Rouge;
37	I2M2=Vert	IBD=Bleu	1	1	0,945	0,111	0,520			0,055	1BILO2_Bleu;
40	I2M2=Vert	IBMR=Rouge	2	2	0,915	0,222	0,160			0,033	2SPD_Rouge;
38	I2M2=Vert	IBMR=Rouge	1	1	0,922	0,111	0,200			0,020	1SPD_Rouge;
17	I2M2=Jaune	IBMR=Rouge	3	3	0,916	0,333	0,200			0,061	3SPD_Rouge;
16	I2M2=Jaune	IBMR=Rouge	2	2	0,923	0,222	0,160			0,033	2SPD_Rouge;
14	I2M2=Jaune	IBMR=Rouge	1	1	0,915	0,111	0,280			0,028	1BILO2_Vert;
15	I2M2=Jaune	IBMR=Rouge	1	1	0,927	0,111	0,200			0,021	1SPD_Rouge;
18	I2M2=Jaune	IBMR=Rouge	4	4	0,903	0,444	0,040			0,016	4SPD_Rouge;
25	I2M2=Orange	IBMR=Rouge	5	5	0,898	0,556	0,520			0,260	5SPD_Rouge;
23	I2M2=Orange	IBMR=Rouge	3	3	0,929	0,333	0,200			0,062	3SPD_Rouge;
22	I2M2=Orange	IBMR=Rouge	2	2	0,930	0,222	0,160			0,033	2SPD_Rouge;
20	I2M2=Orange	IBMR=Rouge	1	1	0,932	0,111	0,280			0,029	1BILO2_Vert;
21	I2M2=Orange	IBMR=Rouge	1	1	0,935	0,111	0,200			0,021	1SPD_Rouge;
34	I2M2=Rouge	IBMR=Rouge	5	5	0,887	0,556	0,920			0,454	1BILO2_Vert;4SPD_Rouge;
32	I2M2=Rouge	IBMR=Rouge	4	4	0,895	0,444	0,920			0,366	1BILO2_Vert;3SPD_Rouge;
35	I2M2=Rouge	IBMR=Rouge	5	5	0,898	0,556	0,520			0,259	5SPD_Rouge;
29	I2M2=Rouge	IBMR=Rouge	2	2	0,901	0,222	0,960			0,192	1BILO2_Vert;1SPD_Rouge;
30	I2M2=Rouge	IBMR=Rouge	2	2	0,882	0,222	0,960			0,188	1BILO2_Vert;1SPD_Rouge;
50	IBD=Bleu	IBD=Bleu	3	3	0,920	0,333	0,800		1,004	1,250	3BILO2_Bleu;
46	IBD=Bleu	IBD=Bleu	2	2	0,956	0,222	0,640		1,003	1,139	2BILO2_Bleu;
44	IBD=Bleu	IBD=Bleu	1	1	0,979	0,111	0,520		1,007	1,063	1BILO2_Bleu;
59	IBD=Bleu	IBMR=Bleu	4	4	0,890	0,444	0,960			0,380	1BILO2_Bleu;3SPD_Rouge;
61	IBD=Bleu	IBMR=Bleu	4	4	0,886	0,444	0,960			0,378	1BILO2_Bleu;3SPD_Rouge;
80	IBD=Vert	IBD=Vert	2	2	0,882	0,222	0,960		1,051	1,239	1BILO2_Bleu;1BILO2_Vert;
87	IBD=Vert	IBMR=Bleu	4	4	0,880	0,444	0,800			0,313	1BILO2_Bleu;3SPD_Rouge;
89	IBD=Vert	IBMR=Rouge	5	5	0,892	0,556	0,520			0,258	5SPD_Rouge;
86	IBD=Vert	IBMR=Bleu	3	3	0,883	0,333	0,800			0,236	1BILO2_Bleu;2SPD_Rouge;
84	IBD=Vert	IBMR=Bleu	2	2	0,886	0,222	0,760			0,150	1BILO2_Bleu;1SPD_Rouge;
66	IBD=Jaune	IBMR=Rouge	3	3	0,904	0,333	0,200			0,060	3SPD_Rouge;
65	IBD=Jaune	IBMR=Rouge	2	2	0,910	0,222	0,160			0,032	2SPD_Rouge;
63	IBD=Jaune	IBMR=Rouge	1	1	0,934	0,111	0,280			0,029	1BILO2_Vert;
64	IBD=Jaune	IBMR=Rouge	1	1	0,912	0,111	0,200			0,020	1SPD_Rouge;
67	IBD=Jaune	IBMR=Rouge	4	4	0,891	0,444	0,040			0,016	4SPD_Rouge;
72	IBD=Orange	IBMR=Rouge	3	3	0,891	0,333	0,200			0,059	3SPD_Rouge;
71	IBD=Orange	IBMR=Rouge	2	2	0,897	0,222	0,160			0,032	2SPD_Rouge;
69	IBD=Orange	IBMR=Rouge	1	1	0,925	0,111	0,280			0,029	1BILO2_Vert;
70	IBD=Orange	IBMR=Rouge	1	1	0,903	0,111	0,200			0,020	1SPD_Rouge;
73	IBD=Orange	IBMR=Rouge	4	4	0,889	0,444	0,040			0,016	4SPD_Rouge;
75	IBD=Rouge	IBD=Bleu	1	1	0,915	0,111	0,520			0,053	1BILO2_Bleu;
93	IBGN=Bleu	IBD=Bleu	2	2	0,914	0,222	0,640			0,130	2BILO2_Bleu;
95	IBGN=Bleu	IBMR=Rouge	3	3	0,906	0,333	0,200			0,060	3SPD_Rouge;
91	IBGN=Bleu	IBD=Bleu	1	1	0,950	0,111	0,520			0,055	1BILO2_Bleu;
94	IBGN=Bleu	IBMR=Rouge	2	2	0,913	0,222	0,160			0,032	2SPD_Rouge;
92	IBGN=Bleu	IBMR=Rouge	1	1	0,920	0,111	0,200			0,020	1SPD_Rouge;
118	IBGN=Vert	IBMR=Rouge	3	3	0,908	0,333	0,200			0,061	3SPD_Rouge;
115	IBGN=Vert	IBD=Bleu	1	1	0,901	0,111	0,520			0,052	1BILO2_Bleu;
117	IBGN=Vert	IBMR=Rouge	2	2	0,916	0,222	0,160			0,033	2SPD_Rouge;
116	IBGN=Vert	IBMR=Rouge	1	1	0,921	0,111	0,200			0,020	1SPD_Rouge;
119	IBGN=Vert	IBMR=Rouge	4	4	0,895	0,444	0,040			0,016	4SPD_Rouge;
101	IBGN=Jaune	IBMR=Rouge	3	3	0,931	0,333	0,200			0,062	3SPD_Rouge;
100	IBGN=Jaune	IBMR=Rouge	2	2	0,938	0,222	0,160			0,033	2SPD_Rouge;
98	IBGN=Jaune	IBMR=Rouge	1	1	0,913	0,111	0,280			0,028	1BILO2_Vert;
99	IBGN=Jaune	IBMR=Rouge	1	1	0,940	0,111	0,200			0,021	1SPD_Rouge;
102	IBGN=Jaune	IBMR=Rouge	4	4	0,912	0,444	0,040			0,016	4SPD_Rouge;
109	IBGN=Orange	IBMR=Rouge	5	5	0,895	0,556	0,520			0,259	5SPD_Rouge;
107	IBGN=Orange	IBMR=Rouge	3	3	0,940	0,333	0,200			0,063	3SPD_Rouge;

n° motif	Contexte de génération	Contexte dominant	Nb items	Nb itemsets	F	C	S	E	P=FCS+E		Items du motif
106	IBGN=Orange	IBMR=Rouge	2	2	0,944	0,222	0,160			0,034	2SPD_Rouge;
104	IBGN=Orange	IBMR=Rouge	1	1	0,907	0,111	0,280			0,028	1BILO2_Vert;
105	IBGN=Orange	IBMR=Rouge	1	1	0,956	0,111	0,200			0,021	1SPD_Rouge;
113	IBGN=Rouge	IBMR=Rouge	5	5	0,900	0,556	0,520			0,260	5SPD_Rouge;
111	IBGN=Rouge	IBMR=Rouge	1	1	0,900	0,111	0,280			0,028	1BILO2_Vert;
112	IBGN=Rouge	IBMR=Rouge	4	4	0,967	0,444	0,040			0,017	4SPD_Rouge;
142	IBMR=Bleu	IBMR=Bleu	6	6	0,903	0,667	0,960		1,053	1,631	2BILO2_Bleu;4SPD_Rouge;
148	IBMR=Bleu	IBMR=Bleu	6	6	0,891	0,667	0,960		1,061	1,631	1BILO2_Bleu;5SPD_Rouge;
147	IBMR=Bleu	IBMR=Bleu	6	6	0,892	0,667	0,960		1,054	1,625	1BILO2_Bleu;5SPD_Rouge;
151	IBMR=Bleu	IBMR=Bleu	6	6	0,893	0,667	0,960		1,048	1,620	2BILO2_Bleu;4SPD_Rouge;
146	IBMR=Bleu	IBMR=Bleu	6	6	0,889	0,667	0,960		1,045	1,614	2BILO2_Bleu;4SPD_Rouge;
185	IBMR=Vert	IBMR=Rouge	5	5	0,893	0,556	0,520			0,258	5SPD_Rouge;
181	IBMR=Vert	IBMR=Bleu	2	2	0,880	0,222	0,760			0,149	1BILO2_Bleu;1SPD_Rouge;
180	IBMR=Vert	IBD=Bleu	2	2	0,892	0,222	0,640			0,127	2BILO2_Bleu;
183	IBMR=Vert	IBMR=Rouge	3	3	0,922	0,333	0,200			0,061	3SPD_Rouge;
178	IBMR=Vert	IBD=Bleu	1	1	0,942	0,111	0,520			0,054	1BILO2_Bleu;
157	IBMR=Jaune	IBMR=Rouge	4	4	0,885	0,444	0,920			0,362	1BILO2_Vert;3SPD_Rouge;
155	IBMR=Jaune	IBMR=Rouge	3	3	0,889	0,333	0,920			0,273	1BILO2_Vert;2SPD_Rouge;
159	IBMR=Jaune	IBMR=Rouge	5	5	0,898	0,556	0,520			0,259	5SPD_Rouge;
156	IBMR=Jaune	IBMR=Rouge	3	3	0,931	0,333	0,200			0,062	3SPD_Rouge;
154	IBMR=Jaune	IBMR=Rouge	2	2	0,934	0,222	0,160			0,033	2SPD_Rouge;
168	IBMR=Orange	IBMR=Rouge	6	6	0,885	0,667	0,920			0,543	6SPD_Rouge;
166	IBMR=Orange	IBMR=Rouge	5	5	0,880	0,556	0,920			0,450	1BILO2_Vert;4SPD_Rouge;
164	IBMR=Orange	IBMR=Rouge	3	3	0,885	0,333	0,920			0,272	1BILO2_Vert;2SPD_Rouge;
167	IBMR=Orange	IBMR=Rouge	5	5	0,901	0,556	0,520			0,260	5SPD_Rouge;
163	IBMR=Orange	IBMR=Rouge	2	2	0,891	0,222	0,960			0,190	2BILO2_Vert;
171	IBMR=Rouge	IBMR=Rouge	9	7	0,905	1,000	0,960		1,279	2,148	2POSPE_Rouge;7SPD_Rouge;
175	IBMR=Rouge	IBMR=Rouge	6	6	0,905	0,667	0,960		1,106	1,686	2BILO2_Vert;4SPD_Rouge;
174	IBMR=Rouge	IBMR=Rouge	6	6	0,905	0,667	0,960		1,086	1,665	1BILO2_Vert;5SPD_Rouge;
172	IBMR=Rouge	IBMR=Rouge	5	5	0,952	0,556	0,960		1,108	1,616	1BILO2_Vert;4SPD_Rouge;
173	IBMR=Rouge	IBMR=Rouge	6	6	0,905	0,667	0,920		1,022	1,577	6SPD_Rouge;
198	IPR=Bleu	IBMR=Bleu	4	4	0,882	0,444	0,800			0,314	1BILO2_Bleu;3SPD_Rouge;
194	IPR=Bleu	IBD=Bleu	3	3	0,900	0,333	0,800			0,240	3BILO2_Bleu;
195	IPR=Bleu	IBMR=Bleu	3	3	0,884	0,333	0,800			0,236	1BILO2_Bleu;2SPD_Rouge;
196	IPR=Bleu	IBMR=Bleu	3	3	0,880	0,333	0,800			0,235	1BILO2_Bleu;2SPD_Rouge;
191	IPR=Bleu	IBMR=Bleu	2	2	0,893	0,222	0,760			0,151	1BILO2_Bleu;1SPD_Rouge;
234	IPR=Vert	IBMR=Bleu	4	4	0,881	0,444	0,800			0,313	1BILO2_Bleu;3SPD_Rouge;
233	IPR=Vert	IBMR=Bleu	3	3	0,881	0,333	0,880			0,258	1BILO2_Bleu;2SPD_Rouge;
230	IPR=Vert	IBMR=Bleu	3	3	0,888	0,333	0,800			0,237	1BILO2_Bleu;2SPD_Rouge;
231	IPR=Vert	IBMR=Bleu	3	3	0,887	0,333	0,800			0,236	1BILO2_Bleu;2SPD_Rouge;
229	IPR=Vert	IBD=Bleu	3	3	0,881	0,333	0,800			0,235	3BILO2_Bleu;
205	IPR=Jaune	IBMR=Rouge	3	3	0,910	0,333	0,200			0,061	3SPD_Rouge;
203	IPR=Jaune	IBD=Bleu	1	1	0,912	0,111	0,520			0,053	1BILO2_Bleu;
204	IPR=Jaune	IBMR=Rouge	2	2	0,916	0,222	0,160			0,033	2SPD_Rouge;
201	IPR=Jaune	IBMR=Rouge	1	1	0,910	0,111	0,280			0,028	1BILO2_Vert;
202	IPR=Jaune	IBMR=Rouge	1	1	0,923	0,111	0,200			0,021	1SPD_Rouge;
213	IPR=Orange	IBMR=Rouge	5	5	0,885	0,556	0,520			0,256	5SPD_Rouge;
211	IPR=Orange	IBMR=Rouge	3	3	0,925	0,333	0,200			0,062	3SPD_Rouge;
210	IPR=Orange	IBMR=Rouge	2	2	0,933	0,222	0,160			0,033	2SPD_Rouge;
209	IPR=Orange	IBMR=Rouge	1	1	0,893	0,111	0,280			0,028	1BILO2_Vert;
208	IPR=Orange	IBMR=Rouge	1	1	0,938	0,111	0,200			0,021	1SPD_Rouge;
220	IPR=Rouge	IBMR=Rouge	5	5	0,911	0,556	0,520			0,263	5SPD_Rouge;
218	IPR=Rouge	IBMR=Rouge	3	3	0,948	0,333	0,200			0,063	3SPD_Rouge;
217	IPR=Rouge	IBMR=Rouge	2	2	0,950	0,222	0,160			0,034	2SPD_Rouge;
215	IPR=Rouge	IBMR=Rouge	1	1	0,890	0,111	0,280			0,028	1BILO2_Vert;
216	IPR=Rouge	IBMR=Rouge	1	1	0,953	0,111	0,200			0,021	1SPD_Rouge;

ANNEXE 6 : Répartition des motifs de l'extraction 2 (HER18, I2M2, 60 mois, fréquence minimale 0,7) et de leur support par classe, triés par combinaison P décroissante

Classe	N° Motif	Nb de Support
1	24	11
1	25	11
1	26	9
1	28	8
1	31	10
1	33	10
2	0	6
2	5	5
2	6	5
2	7	3
2	8	2
2	10	3
2	11	2
3	27	10
3	29	9
3	30	10
3	32	9
3	34	9
3	35	10
3	36	9
4	172	6
4	173	5
4	174	6
4	176	4
4	177	6
4	178	5
4	180	6
5	19	5
5	179	23
5	181	23
5	182	23
6	3	6
6	23	5

Classe	N° Motif	Nb de Support
7	109	5
7	110	5
7	111	4
7	112	4
7	113	4
7	114	5
7	115	5
7	116	2
7	117	3
7	118	2
7	119	4
7	120	3
7	121	4
7	122	2
7	123	3
7	124	2
7	125	4
7	126	1
7	127	4
7	128	3
7	129	4
7	130	3
7	131	2
7	132	3
7	133	3
7	134	2
7	135	4
7	136	3
7	137	3
7	138	3
7	139	3
7	140	4
7	141	2
7	142	4
7	143	3
7	144	2
7	145	3
7	146	5
7	147	2
7	148	5
7	149	5
7	150	4
7	151	5
7	152	5

Classe	N° Motif	Nb de Support
8	37	14
8	40	13
8	41	14
8	42	8
8	43	14
8	45	12
8	46	12
8	50	12
8	51	11
8	52	12
8	53	13
8	54	13
8	55	11
8	56	13
8	57	12
8	58	10
8	59	11
8	60	11
8	61	12
8	62	11
8	63	11
8	64	12
8	66	13
8	67	13
8	68	12
8	71	13
8	73	11
8	74	12
8	75	11
8	76	11
8	77	11
8	80	10
8	81	12
8	83	12
8	84	13
8	86	11
8	87	10
8	90	13
8	91	11
8	92	10
8	95	10
8	96	11
8	100	11
8	101	12

Classe	N° Motif	Nb de Support
9	38	9
9	39	12
9	44	9
9	47	9
9	48	9
9	49	8
9	78	12
9	107	12
10	13	5
10	14	4
10	16	5
10	17	4
10	20	3
10	21	4
11	4	6
11	12	5
12	2	6
12	9	5
13	1	4
13	15	5
13	18	4
13	22	4
14	65	12
14	69	11
14	70	14
14	72	12
14	79	13
14	82	11
14	85	13
14	88	13
14	89	11
14	93	11
14	94	12
14	97	10
14	98	10
14	99	12
14	104	12
14	105	11
14	108	11

Classe	N° Motif	Nb de Support	Classe	N° Motif	Nb de Support	Classe	N° Motif	Nb de Support	Classe	N° Motif	Nb de Support
			7	153	5	8	102	12			
			7	154	2	8	103	11			
			7	155	3	8	106	11			
			7	156	4	8	175	24			
			7	157	5						
			7	158	4						
			7	159	4						
			7	160	4						
			7	161	4						
			7	162	4						
			7	163	3						
			7	164	3						
			7	165	4						
			7	166	4						
			7	167	3						
			7	168	4						
			7	169	3						
			7	170	3						
			7	171	4						

ANNEXE 7 : Motifs sélectionnés par classe, dont le support est égal à la médiane des supports, et triés par combinaison P=(FxCxS + E) décroissant (extraction HER18, I2M2, 60 mois, fréquence minimale 0,7)

Classe	Motif	Nb supports	Contexte de génération	Contexte dominant	F	C	S	E	P	Items du motif
1	33	10	I2M2=Jaune	I2M2=Bleu	0,778	0,012	0,800		0,008	3 NITR_Vert
1	31	10	I2M2=Jaune	I2M2=Bleu	0,806	0,008	0,800		0,005	2 NITR_Vert
2	10	3	I2M2=Bleu	I2M2=Bleu	0,765	0,049	0,800	2,676	2,706	10 NITR_Vert, 2 MINE_Jaune
2	7	3	I2M2=Bleu	I2M2=Bleu	0,824	0,033	0,800	2,661	2,682	5 NITR_Vert, 3 MINE_Jaune
3	34	9	I2M2=Jaune	I2M2=Bleu	0,750	0,016	0,800		0,010	2NITR_Vert, 2 PAES_Jaune
3	36	9	I2M2=Jaune	I2M2=Bleu	0,722	0,016	0,800		0,009	3 NITR_Vert, 1 PHOS_Jaune
3	29	9	I2M2=Jaune	I2M2=Rouge	0,750	0,008	0,800		0,005	1 NITR_Jaune, 1 PAES_Jaune
3	32	9	I2M2=Jaune	I2M2=Rouge	0,722	0,008	0,800		0,005	1 NITR_Jaune, 1 PHOS_Jaune
4	180	6	I2M2=Vert	I2M2=Rouge	0,738	0,012	0,800		0,007	3 PHOS_Jaune
4	177	6	I2M2=Vert	I2M2=Orange	0,786	0,008	0,600		0,004	2 PHOS_Jaune
4	174	6	I2M2=Vert	I2M2=Rouge	0,857	0,004	0,600		0,002	1 PHOS_Jaune
4	172	6	I2M2=Vert	I2M2=Bleu	1,000	0,000	0,800		0,000	vide
5	182	23	I2M2=Vert	I2M2=Bleu	0,714	0,029	0,800		0,016	7 NITR_Vert
5	181	23	I2M2=Vert	I2M2=Bleu	0,738	0,020	0,800		0,012	5 NITR_Vert
5	179	23	I2M2=Vert	I2M2=Bleu	0,762	0,016	0,800		0,010	4 NITR_Vert
6	23	5	I2M2=Bleu	I2M2=Bleu	0,706	0,151	0,800	2,281	2,366	29 NITR_Vert, 9 MINE_Vert
6	3	6	I2M2=Bleu	I2M2=Rouge	0,706	0,008	0,800		0,005	1 MOOX_Jaune, 1 PHOS_Jaune
7	165	4	I2M2=Rouge	I2M2=Rouge	0,786	0,567	0,800	11,786	12,142	8 MINE_Vert, 15 NITR_Jaune, 13 NITR_Orange, 43 PHOS_Jaune, 17 AZOT_Jaune, 27 MOOX_Jaune, 12 PAES_Jaune, 4 PAES_Rouge
7	171	4	I2M2=Rouge	I2M2=Rouge	0,714	0,976	0,800	10,714	11,272	10 MINE_Vert, 20 NITR_Jaune, 28 NITR_Orange, 77 PHOS_Jaune, 27 AZOT_Jaune, 48 MOOX_Jaune, 26 PAES_Jaune, 9 PAES_Rouge
7	168	4	I2M2=Rouge	I2M2=Rouge	0,714	0,645	0,600	10,714	10,991	14 MINE_Vert, 26 NITR_Jaune, 12 NITR_Orange, 48 PHOS_Jaune, 22 AZOT_Jaune, 23 MOOX_Jaune, 20 PAES_Jaune, 5 PAES_Rouge
7	166	4	I2M2=Rouge	I2M2=Rouge	0,714	0,465	0,600	10,714	10,914	8 MINE_Vert, 12 NITR_Jaune, 11 NITR_Orange, 36 PHOS_Jaune, 16 AZOT_Jaune, 20 MOOX_Jaune, 9 PAES_Jaune, 3 PAES_Rouge
7	129	4	I2M2=Rouge	I2M2=Rouge	0,714	0,143	0,800	7,143	7,224	3 MINE_Vert, 6 NITR_Jaune, 16 PHOS_Jaune, 5 AZOT_Jaune, 5 PAES_Jaune

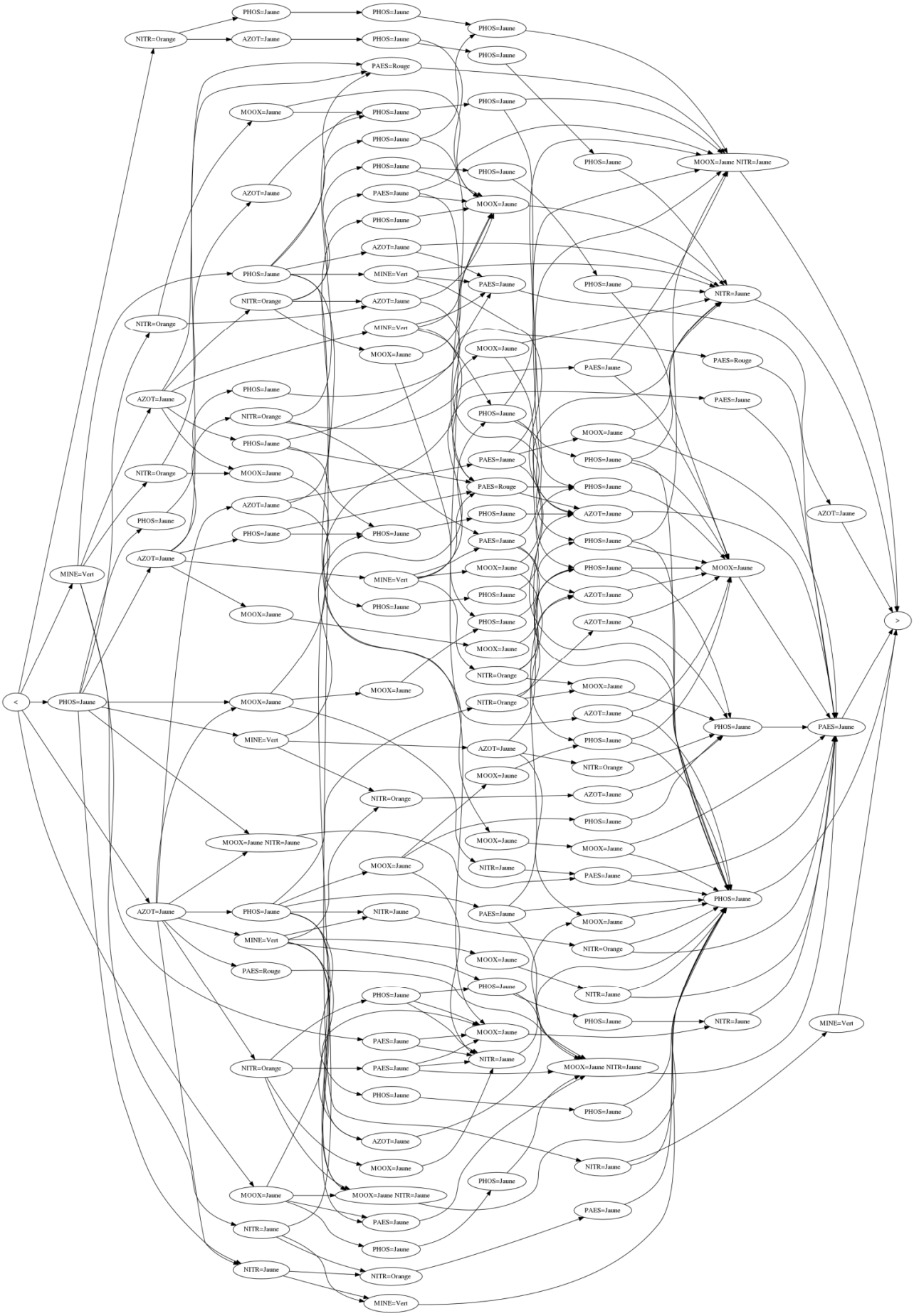
Classe	Motif	Nb supports	Contexte de génération	Contexte dominant	F	C	S	E	P	Items du motif
7	121	4	I2M2=Rouge	I2M2=Rouge	0,857	0,053	0,800	6,429	6,465	1 MINE_Vert, 2 NITR_Orange, 3 PHOS_Jaune, 2 AZOT_Jaune, 2 MOOX_Jaune, 3 PAES_Rouge
7	125	4	I2M2=Rouge	I2M2=Rouge	0,786	0,053	0,800	5,893	5,926	1 MINE_Vert, 4 NITR_Orange, 3 PHOS_Jaune, 2 AZOT_Jaune, 1 MOOX_Jaune, 1 PAES_Rouge
7	159	4	I2M2=Rouge	I2M2=Rouge	0,714	0,200	0,800	5,357	5,471	4 MINE_Vert, 13 NITR_Jaune, 23 PHOS_Jaune, 4 AZOT_Jaune,5 PAES_Jaune
7	127	4	I2M2=Rouge	I2M2=Rouge	0,786	0,057	0,800	4,714	4,750	2 MINE_Vert, 2 NITR_Jaune, 2 PHOS_Jaune, 5 AZOT_Jaune, 3MOOX_Jaune
7	162	4	I2M2=Rouge	I2M2=Rouge	0,714	0,220	0,800	4,286	4,412	6 MINE_Vert, 12 NITR_Jaune, 26 PHOS_Jaune, 4 AZOT_Jaune, 4 PAES_Jaune
7	156	4	I2M2=Rouge	I2M2=Rouge	0,714	0,131	0,800	4,286	4,360	5 MINE_Vert, 4 NITR_Jaune, 12 PHOS_Jaune, 7 AZOT_Jaune, 4 PAES_Jaune
7	119	4	I2M2=Rouge	I2M2=Rouge	0,857	0,053	0,800	4,286	4,322	1 MINE_Vert, 2 NITR_Jaune, 2 PHOS_Jaune, 5 AZOT_Jaune, 3 MOOX_Jaune
7	161	4	I2M2=Rouge	I2M2=Rouge	0,786	0,208	0,800	3,929	4,059	4 MINE_Vert, 12 NITR_Jaune, 25 PHOS_Jaune, 5 AZOT_Jaune, 5 PAES_Jaune
7	150	4	I2M2=Rouge	I2M2=Rouge	0,786	0,118	0,800	3,929	4,003	2 MINE_Vert, 5 NITR_Jaune, 12 PHOS_Jaune, 6 AZOT_Jaune, 4 PAES_Jaune
7	160	4	I2M2=Rouge	I2M2=Rouge	0,714	0,192	0,800	3,571	3,681	4 MINE_Vert, 11 NITR_Jaune, 25 PHOS_Jaune, 4 AZOT_Jaune, 4 PAES_Jaune
7	158	4	I2M2=Rouge	I2M2=Rouge	0,714	0,114	0,800	3,571	3,637	3 MINE_Vert, 4 NITR_Jaune, 12 PHOS_Jaune, 5 AZOT_Jaune, 4 PAES_Jaune
7	142	4	I2M2=Rouge	I2M2=Rouge	0,857	0,037	0,800	3,214	3,239	1 MINE_Vert, 1 NITR_Jaune, 3 PHOS_Jaune, 4 AZOT_Jaune
7	140	4	I2M2=Rouge	I2M2=Rouge	0,857	0,029	0,800	2,857	2,877	1 MINE_Vert, 2 NITR_Jaune, 2 PHOS_Jaune, 3 AZOT_Jaune
7	112	4	I2M2=Rouge	I2M2=Rouge	0,929	0,016	0,800	2,786	2,798	1 MINE_Vert, 2 AZOT_Jaune, 1 PHOS_Jaune
7	113	4	I2M2=Rouge	I2M2=Rouge	0,786	0,012	0,800	2,619	2,627	1 MINE_Vert, 1 AZOT_Jaune, 1 PHOS_Jaune
7	135	4	I2M2=Rouge	I2M2=Rouge	0,857	0,020	0,800	2,571	2,585	1 MINE_Vert, 2 AZOT_Jaune, 2 PHOS_Jaune
7	111	4	I2M2=Rouge	I2M2=Rouge	0,929	0,020	0,800	2,321	2,337	1 MINE_Vert, 2 AZOT_Jaune, 1 PHOS_Jaune, 1 MOOX_Jaune
8	74	12	I2M2=Orange	I2M2=Orange	0,700	0,020	0,800	1,225	1,236	2 NITR_Vert, 2 PHOS_Jaune, 1PAES_Rouge
8	101	12	I2M2=Orange	I2M2=Orange	0,700	0,045	0,800	1,089	1,114	3 NITR_Jaune, 5 PHOS_Jaune, 3 MOOX_Jaune
8	83	12	I2M2=Orange	I2M2=Orange	0,700	0,041	0,800	1,089	1,112	4 NITR_Jaune, 5 PHOS_Jaune, 1 MOOX_Jaune
8	102	12	I2M2=Orange	I2M2=Orange	0,767	0,037	0,800	1,073	1,096	3 NITR_Jaune, 5 PHOS_Jaune, 1 MOOX_Jaune
8	50	12	I2M2=Orange	I2M2=Orange	0,733	0,012	0,800	1,027	1,034	3 NITR_Jaune
8	57	12	I2M2=Orange	I2M2=Orange	0,800	0,016	0,800	1,018	1,029	2 NITR_Jaune, 2 PHOS_Jaune
8	61	12	I2M2=Orange	I2M2=Orange	0,700	0,012	0,800	1,008	1,015	1 NITR_Vert, 2 PHOS_Jaune
8	64	12	I2M2=Orange	I2M2=Rouge	0,700	0,016	0,800		0,009	4 NITR_Jaune

Classe	Motif	Nb supports	Contexte de génération	Contexte dominant	F	C	S	E	P	Items du motif
8	52	12	I2M2=Orange	I2M2=Rouge	0,767	0,012	0,800		0,008	1 PHOS_Jaune, 1 MOOX_Jaune, 1PAES_Rouge
8	68	12	I2M2=Orange	I2M2=Rouge	0,733	0,012	0,800		0,007	1 PHOS_Jaune, 1 MOOX_Jaune, 1PAES_Rouge
8	46	12	I2M2=Orange	I2M2=Rouge	0,833	0,008	0,800		0,005	2 NITR_Jaune
8	45	12	I2M2=Orange	I2M2=Rouge	0,800	0,008	0,800		0,005	1 PHOS_Jaune, 1PAES_Rouge
8	81	12	I2M2=Orange	I2M2=Rouge	0,767	0,008	0,800		0,005	1 PHOS_Jaune, 1PAES_Rouge
9	48	9	I2M2=Orange	I2M2=Orange	0,700	0,008	0,800	1,633	1,638	1 MOOX Bleu, 1 PHOS_Jaune
9	44	9	I2M2=Orange	I2M2=Orange	0,733	0,008	0,800	1,467	1,471	1 MOOX Bleu, 1 PHOS_Jaune
9	47	9	I2M2=Orange	I2M2=Orange	0,700	0,008	0,800	1,400	1,405	1 MOOX Bleu, 1 NITR_Jaune
9	38	9	I2M2=Orange	I2M2=Orange	0,767	0,004	0,800	1,400	1,403	1 MOOX Bleu
10	21	4	I2M2=Bleu	I2M2=Bleu	0,706	0,143	0,800	2,471	2,551	24 NITR_Vert, 11 MINE_Vert
10	17	4	I2M2=Bleu	I2M2=Bleu	0,706	0,098	0,800	2,281	2,336	17 NITR_Vert, 7 MINE_Vert
10	14	4	I2M2=Bleu	I2M2=Bleu	0,765	0,078	0,800	1,784	1,832	15 NITR_Vert, 4 MINE_Vert
11	12	5	I2M2=Bleu	I2M2=Bleu	0,706	0,049	0,800	1,412	1,439	10 NITR_Vert, 2 PHOS_Jaune
11	4	6	I2M2=Bleu	I2M2=Orange	0,706	0,008	0,600		0,003	2 PHOS_Jaune
12	9	5	I2M2=Bleu	I2M2=Bleu	0,765	0,029	0,800	1,434	1,451	6 NITR_Vert, 1 PHOS_Jaune
12	2	6	I2M2=Bleu	I2M2=Rouge	0,824	0,004	0,600		0,002	1 PHOS_Jaune
13	22	4	I2M2=Bleu	I2M2=Bleu	0,706	0,090	0,800	2,281	2,331	20 NITR_Vert, 1MINE_Vert
13	18	4	I2M2=Bleu	I2M2=Bleu	0,706	0,082	0,800	1,976	2,023	15 NITR_Vert, 4 MINE_Vert
13	1	4	I2M2=Bleu	I2M2=Rouge	0,706	0,004	0,800		0,002	1 NITR_Jaune
14	94	12	I2M2=Orange	I2M2=Orange	0,700	0,029	0,800	1,089	1,105	3 PHOS_Jaune, 4 MOOX_Jaune
14	104	12	I2M2=Orange	I2M2=Orange	0,733	0,041	0,800	1,027	1,051	4 NITR_Jaune, 5 PHOS_Jaune, 1 MOOX_Jaune
14	99	12	I2M2=Orange	I2M2=Rouge	0,700	0,037	0,800		0,021	4 PHOS_Jaune, 5 MOOX_Jaune
14	65	12	I2M2=Orange	I2M2=Rouge	0,700	0,016	0,800		0,009	2 PHOS_Jaune, 1 MOOX_Jaune, 1 PAES_Rouge
14	79	13	I2M2=Orange	I2M2=Rouge	0,700	0,016	0,800		0,009	2 NITR_Jaune, 1 PHOS_Jaune, 1 MOOX_Jaune
14	72	12	I2M2=Orange	I2M2=Rouge	0,733	0,012	0,800		0,007	1 PHOS_Jaune, 1 MOOX_Jaune, 1 PAES_Rouge

ANNEXE 8 : Nombre de stations réparties par type d'état biologique et vérifiant les classes de motifs obtenues pour l'extraction 2 (extraction HER18, I2M2, 60 mois, fréquence minimale 0,7)

Classe des motifs	Types de l'état biologique I2M2 des stations				
	S(1-3)	V(+)	S(3-5)	V(-)	TV
1	16	35	22	24	4
2	14	16	0	3	0
3	15	40	26	28	4
4	17	47	7	27	6
5	8	21	3	13	0
6	4	7	0	2	0
7	0	0	125	115	0
8	2	35	209	131	32
9	0	9	28	24	7
10	12	15	0	4	0
11	4	7	0	2	0
12	4	7	0	2	0
13	8	11	0	3	0
14	0	7	77	49	11
Métriques statistiques par type de stations					
Médiane	*6*	*13*	*5*	*18,5*	*0*
1°quartile	*2,5*	*7*	*0*	*3*	*0*
3°quartile	*14*	*32*	*28*	*28*	*6*

ANNEXE 9 : Motif 165 (extraction HER18, I2M2, 60 mois, fréquence minimum 0,7)



Université de Strasbourg | ENGEES | Image Ville Environnement | ED 413

Corinne GRAC

Fouille temporelle des indicateurs physico-chimiques et biologiques pour l'évaluation de l'état, des pressions et de la capacité de résilience des rivières

ANNEXES

RESUME

Les données issues de la surveillance des rivières sont volumineuses, avec des relations complexes. Des méthodes de fouille de données non supervisées peuvent s'y appliquer et donner des résultats pertinents pour leur gestion, sous réserve d'une collaboration étroite entre hydroécologues et informaticiens. L'extraction de motifs partiellement ordonnés à partir de séquences temporelles de pressions physico-chimiques précédant un état biologique a été réalisée. Ces motifs temporels permettent d'identifier une partie des pressions en cause dans un état écologique dégradé ou non, de préciser l'importance de la durée des séquences avant l'évaluation de l'état biologique, d'identifier les altérations caractéristiques à l'échelle d'une hydro-écorégion. A terme nous envisageons d'élargir ces motifs aux pressions hydromorphologiques.

MOTS CLES: rivière, pressions physico-chimiques, indices biologiques, macro-invertébrés, poissons, diatomées, macrophytes, état écologique, Directive Cadre européenne sur l'Eau, pluridisciplinarité, fouille de données, motifs temporels

ABSTRACT

Data from the assessment of river are big data, with complex relationships. Unsupervised data mining methods can be applied on them and give relevant results for their management, if a close collaboration exists between hydroecologists and computer scientists. The extraction of partially ordered patterns from temporal sequences of physicochemical pressures preceding a biological state has been achieved. These temporal patterns allow to identify a part of the pressures involved or not in a degraded ecological status, to specify the importance of the sequences time-length before a biological assessment, to identify the characteristic pressure categories at a regional scale. To go further, we plan to extend these patterns to hydromorphological pressures.

KEYWORDS: Rivers, physico-chemical pressures, biological indices, macroinvertebrates, fish, diatoms, macrophytes, ecological status, European Water Framework Directive, pluridisciplinarity, data mining, temporal patterns