Recherche de motifs partiellement ordonnés clos discriminants pour caractériser l'état des milieux aquatiques

Mickaël Fabrègue, Agnès Braud, Sandra Bringay, Florence Le Ber, Maguelonne Teisseire

Recherche de motifs partiellement ordonnés clos discriminants pour caractériser l'état des milieux aquatiques

M. Fabrègue[1,2] A. Braud[1] S. Bringay[3] F. Le Ber[1] M. Teisseire[2]

[1] ICube, Université de Strasbourg, ENGEES, CNRS
[2] IRSTEA, TETIS
[3] LIRMM, Université Montpellier 2

ICUBE, 300 bd Sébastien Brant - CS 10413 - F-67412 Illkirch Cedex
agnes.braud@unistra.fr, florence.leber@engees.unistra.fr
TETIS - 500 rue Jean-François Breton - F-34093 Montpellier Cedex 5
mickael.fabregue@teledetection.fr, maguelonne.teisseire@teledetection.fr
LIRMM - 161 rue Ada - CC477 - F-34095 Montpellier Cedex 5
sandra.bringay@lirmm.fr

Résumé

Cet article présente un processus de fouille de données mis en œuvre pour extraire des connaissances d'un jeu de données concernant l'hydro-écologie des cours d'eau. L'approche s'appuie sur la recherche de motifs clos partiellement ordonnés, utilisés comme éléments discriminants pour relier les paramètres physico-chimiques et biologiques mesurés sur des stations de rivières. Pour chaque valeur d'un indice biologique, sont mis ainsi en évidence des séquences temporelles de valeurs de paramètres physico-chimiques ayant un impact sur la biologie. L'approche est mise en œuvre sur un jeu de données regroupant plusieurs milliers de stations de rivières.

Mots Clef

Fouille de données, motifs clos partiellement ordonnés, mesure d'intérêt, hydro-écologie.

Abstract

This paper presents a data mining process implemented to extract original knowledge from hydro-ecological data. The approach is based on closed partially ordered patterns used as discriminant features to link physico-chemistry with biology in river sampling sites. For each bio-indicator quality value, we obtain a set of significant discriminant features. We use them to identify the impact of physico-chemical characteristics on the biological dimensions. The approach has been experimented on a dataset of several thousands river sites.

Keywords

Data mining, Closed partially ordered patterns, Interestingness measure, Hydro-ecology.

1 Introduction

Le travail présenté ici se situe dans le cadre du projet Fresqueau [1], qui a pour but de développer des méthodes de fouille de données pour étudier, comparer et exploiter l'ensemble des données disponibles permettant d'évaluer l'état d'un cours d'eau. Ces données sont relatives à la qualité de l'eau, l'hydrologie, les stations de mesure, etc. mais concernent également l'environnement des cours d'eau. Il s'agit en particulier des données physico-chimiques et biologiques produites par les agences de l'eau et l'ONEMA (Office National de l'Eau et des Milieux Aquatiques), qui se déclinent en trois sous-ensembles :

1. données concernant l'état physico-chimique de l'eau et des sédiments ; macropolluants (nitrates, matières organiques, ...) et micropolluants (pesticides, ...) ;
2. données concernant l'état des peuplements biologiques floristiques et faunistiques : cet état est synthétisé dans des indices biologiques, parmi lesquels l'indice biologique global normalisé (IBGN) [1] est le plus fréquemment utilisé ;
3. données concernant l'état physique : il s'agit de l'hydromorphologie du cours d'eau (état des berges, du lit mineur, du lit majeur, ...) et des conditions hydrologiques (débits) et hydrauliques (vitesse, géométrie du cours d'eau).

Ces données sont issues de résultats d'analyse de prélèvements effectués régulièrement sur les réseaux de mesure nationaux. Chaque station d'un réseau de mesure est ainsi caractérisée théoriquement par une note annuelle d'un ou plusieurs indices biologiques, et par des valeurs bimensuelles de paramètres physico-chimiques. Dans le cadre de

1. http://engees-fresqueau.unistra.fr/

Numéro Station	Mois / année	NH_4^+	NKJ	NO_2^-	PO_4^{3-}	Phosphore total	IBGN
1	02/07	-	-	-	0,043	0,032	-
	06/07	-	0,672	0,026	-	-	-
	07/07	0,088	1,235	0,134		0,011	-
	09/07	-	-	-	-	-	17
	12/07	0,154	-	0,246	0,168	0,006	-
	02/08	0,062	0,040	0,091	0,025	0,003	-
	04/08	-	0,023	0,198	-	-	-
	05/08	-	-	-	-	-	12
	07/08	-	-	-	0,046	0,009	-
2	01/04	0,043	0,146	0,421	-	-	-
	04/04	-	-	-	0,325	0,093	-
	07/04	2,331	7,993	0,252	0,132	0,066	-
	08/04	-	1,414	-	-	-	-
	09/04	-	-	-	-	-	8
	11/04	0,117	0,0844	-	0,188	-	-
	12/04	-	-	-	0,067	0,078	-
	03/05	-	0,182	0,0310	0,137	-	-
	06/05	0,004	-	0,012	0,035	0,034	-
	08/05	-	-	-	-	-	10

TABLE 1 – Extrait du jeu de données avec différents paramètres physico-chimiques (ammonium, nitrate de Kjeldahl, nitrite, orthophosphate, phosphore total) et un indice biologique (IBGN)

Fresqueau, nous travaillons sur les districts Rhin-Meuse et Rhône-Méditerranée et Corse, qui recouvrent respectivement 33.000 et 130.000 km^2 et comptent ensemble 11.329 stations. Les données sont disponibles sur une dizaine d'années (2000-2010). Pour plus d'information sur les données collectées, le lecteur pourra se référer à [4].
Nous avons exploité ces données pour mettre en évidence des liens entre différentes métriques permettant de caractériser la qualité des cours d'eau, ici les métriques liées aux paramètres physico-chimiques et les métriques liées à la biologie. Nous avons en particulier cherché à mettre en évidence les valeurs des paramètres physico-chimiques précédant généralement dans le temps les valeurs des indices biologiques, en particulier l'IBGN. Notre démarche s'inscrit dans le domaine général de la fouille de motifs dans des séquences, domaine pour lequel de nombreuses méthodes et applications ont été développées [2, 12, 6, 15]. L'article présente brièvement les données, puis la méthode développée spécifiquement pour notre application et quelques résultats obtenus, avant de conclure.

2 Matériel et méthodes

Nous avons exploité les données en utilisant une méthode décrite dans [11], qui recherche des motifs partiellement ordonnés clos (CPO-motifs) dans une base de séquences. Les CPO-motifs ont fait l'objet de différents travaux [14, 5]. Ils présentent un certain nombre d'avantages :

1. ils sont bien adaptés au caractère temporel des données ;
2. ils donnent plus d'information sur l'ordre entre éléments que les motifs séquentiels ;
3. ils sont représentés sous la forme de graphes acycliques, qui sont facilement lisibles par des experts.

Le fait que les motifs soient clos permet de résumer au mieux les informations contenues dans une base de séquences. Malgré cela, la méthode produit encore trop de résultats et pour les réduire, nous utilisons des mesures d'intérêt [12] qui permettent alors de mettre en évidence des motifs discriminants que nous noterons DCPO-motifs par la suite.

2.1 Préparation des données

Avant de mettre en œuvre cette méthode, différents prétraitements sont nécessaires. Nous utilisons l'exemple présenté dans le tableau 1 pour les expliquer. On remarque sur ce tableau que les données sont éparses : en particulier, alors que les paramètres physico-chimiques sont mesurés tous les deux mois par les agences de l'eau, l'IBGN est réalisé au mieux une fois l'an, en période d'étiage, donc généralement en été.
Les données ont d'abord été discrétisées en utilisant la norme SEQ-Eau [2], modifiée à la marge selon l'avis des hydro-écologues travaillant dans le projet Fresqueau. Cette discrétisation transforme les données initiales en cinq classes de qualité, "Très bon", "Bon", "Moyen", "Mauvais" and "Très mauvais" representées par cinq couleurs *Bleu*, *Vert*, *Jaune*, *Orange* et *Rouge*. Par exemple PO_4^{3-}=0,026 appartient à la classe de qualité "Très bon"/*Bleu* et phos-

2. http://sierm.eaurmc.fr/eaux-superficielles/fichiers-telechargeables/grilles-seq-eau-v2.pdf

Sous-bases	Séquences
IBGN^B	$\langle$(AZOT^B)(AZOT^V,PHOS^B)$\rangle$ $\langle$(AZOT^B,PHOS^V)(PHOS^V)(AZOT^J,PHOS^B)$\rangle$
IBGN^J	$\langle$(AZOT^V,PHOS^V)(AZOT^B,PHOS^V)$\rangle$ $\langle$(PHOS^O)(AZOT^O,PHOS^J)(AZOT^V,PHOS^J)$\rangle$ $\langle$(AZOT^V,PHOS^J)(AZOT^V,PHOS^B)(AZOT^V)$\rangle$
IBGN^R	$\langle$(AZOT^O,PHOS^J)(AZOT^R,PHOS^O)(AZOT^V,PHOS^J)$\rangle$ $\langle$(PHOS^O)(AZOT^O,PHOS^J)(AZOT^V)$\rangle$

TABLE 2 – Sous-bases de séquences associées à trois valeurs de l'indice IBGN

phore total = 0,67 appartient à la classe de qualité "Moyen"/*Jaune*. La norme SEQ-Eau permet également de regrouper les paramètres initiaux en 15 macro-paramètres : ainsi les paramètres PO_4^{3-} et phosphore total sont rassemblés en un seul paramètre qualitatif PHOS (matières phosphorées) qui prend la valeur la plus basse des deux ("Moyen"/*Jaune* sur l'exemple précédent). Les paramètres NH_4^+, NJK et NO_2^- sont eux rassemblés dans le macro-paramètre AZOT (matières azotées hors nitrate).
A la suite de cette discrétisation, le jeu de données est transformé en séquences. Chaque valeur de paramètre constitue un *item*. Deux paramètres (ou plus) mesurés à la même date constituent un ensemble d'*items* ou *itemset*. Une succession d'*itemsets* constitue une séquence. Sur la station 1, par exemple, on a relevé la classe de qualité *Vert* pour le macro-paramètre PHOS en février 2007, la classe de qualité *Bleu* pour le macro-paramètre AZOT en juin 2007, la classe de qualité *Vert* pour AZOT et la classe de qualité *Bleu* pour PHOS en juillet 2007. L'indice IBGN mesuré en septembre de la même année a la valeur *Bleu*. Ainsi la station 1 est associée à la séquence $\langle$(PHOS^V) (AZOT^B) (AZOT^V, PHOS^B) (IBGN^B)$\rangle$.
Les séquences ainsi constituées sont ensuite segmentées selon des fenêtres de longueurs 2 à 6 mois, en amont d'une mesure de l'indice biologique. En effet, l'indice biologique intègre les événements physico-chimiques antérieurs sur une durée limitée, et c'est ce que nous cherchons à mettre en évidence. La base de séquences obtenue est alors subdivisée selon la valeur de l'IBGN concluant une séquence. Dans le tableau 2, trois valeurs d'indices sont considérées, *Bleu*, *Jaune* et *Rouge*.

2.2 Extraction des motifs

Cette étape est le cœur de notre proposition. Pour extraire les motifs des différentes sous-bases, nous avons utilisé une version adaptée de l'algorithme présenté dans [11]. La modification essentielle concerne l'utilisation de différents supports minimums, en fonction des sous-bases. Par exemple le motif représenté sur la figure 1 a un support de 0/2 dans la sous-base IBGN^B, 1/3 dans la sous-base IBGN^J et 2/2 dans la sous-base IBGN^R (voir tableau 2).
De plus, nous ne considérons que les motifs partiellement ordonnés clos (CPO-motifs), ce qui permet d'extraire un ensemble plus petit de motifs sans perte d'information. Ainsi le motif P de la figure 1 ne sera pas retenu car un motif l'englobant (présenté sur la figure 2) existe avec le même support. En effet l'*item* PHOS^J est présent dans toutes les séquences supportant P dans les sous-bases IBGN^J et IBGN^R (voir tableau 2).

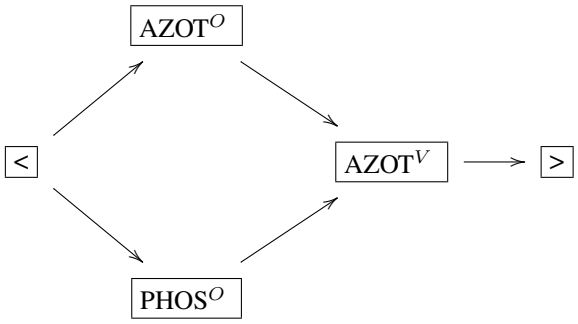

FIGURE 1 – Un exemple de motif partiellement ordonné

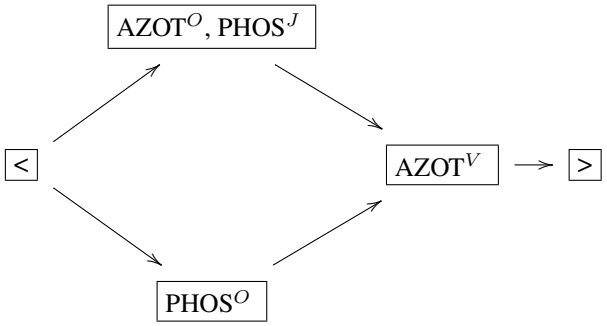

FIGURE 2 – Un exemple de CPO-motif incluant le motif de la figure 1

2.3 Sélection des motifs

Cette sélection s'effectue pour chaque sous-base selon trois critères : la fréquence, la discriminance et la redondance. Fréquence et discriminance sont des critères utilisés par l'algorithme pour limiter l'exploration, à l'issue de laquelle un sous-ensemble de motifs non redondants est sélectionné, selon une procédure définie ci-desous, pour être présenté à l'analyste.

Fréquence d'un CPO-motif : elle est classiquement mesurée par le nombre de séquences de la sous-base qui contiennent le CPO-motif, rapporté au nombre total de séquences dans la sous-base. Lors de l'extraction des CPO-motifs, l'utilisation d'un support minimal permet de sélectionner les motifs fréquents.

Discriminance d'un CPO-motif : elle est mesurée par un facteur de croissance généralisé à partir de celui proposé dans [9]. Un motif est discriminant pour une sous-base s'il est plus fréquent dans cette sous-base que dans toutes les autres. Plus formellement :

Définition 1 *Soit un motif partiellement ordonné clos P, une sous-base $\mathcal{C}$ et un ensemble de sous-bases $\mathcal{E}_\mathcal{C}$ =$\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n\}$, le taux de croissance généralisé de P dans $\mathcal{C}$ en référence à $\mathcal{E}_\mathcal{C}$, dénoté GGR($P, \mathcal{C}, \mathcal{E}_\mathcal{C}$), est :*

$$\begin{cases} 0, \textit{ si } \mathtt{supp}_\mathcal{C}(P) = 0 \textit{ et max}(\mathtt{supp}_{\mathcal{C}_1}(P), \\ \mathtt{supp}_{\mathcal{C}_2}(P), \ldots, \mathtt{supp}_{\mathcal{C}_n}(P)) = 0 \\ \infty, \textit{ si } \mathtt{supp}_\mathcal{C}(P) \neq 0 \textit{ et max}(\mathtt{supp}_{\mathcal{C}_1}(P), \\ \mathtt{supp}_{\mathcal{C}_2}(P), \ldots, \mathtt{supp}_{\mathcal{C}_n}(P)) = 0 \\ \frac{\mathtt{supp}_\mathcal{C}(P)}{max(\mathtt{supp}_{\mathcal{C}_1}(P), \mathtt{supp}_{\mathcal{C}_2}(P), \ldots, \mathtt{supp}_{\mathcal{C}_n}(P))} \textit{ sinon} \end{cases} \quad (1)$$

où $\mathtt{supp}_{\mathcal{C}_i}(P)$ dénote le support du motif P dans la sous-base $\mathcal{C}_i$.

Un CPO-motif est discriminant si GGR($P, \mathcal{C}, \mathcal{E}_\mathcal{C}$)$> 1$. Il est alors noté DCPO-motif pour "motif partiellement ordonné clos discriminant". Le calcul de la discriminance est fait au cours de l'extraction : pour chaque CPO-motif P extrait de la sous-base courante, on calcule son support dans les autres bases, ce qui permet de calculer en même temps le taux de croissance généralisé de P.

Redondance d'un CPO-motif : elle est mesurée par le nombre de CPO-motifs décrivant le même ensemble de séquences. Plus précisément, à chaque CPO-motif est associé un sous-ensemble des séquences de la base, représenté comme une suite binaire de présence-absence de chaque séquence. On peut ainsi utiliser une distance de Hamming entre les ensembles associés à différents CPO-motifs : plus les motifs sont proches (donc couvrent approximativement le même sous-ensemble de séquences) plus ils sont redondants.

Combinaison des critères : les critères de fréquence, discriminance et redondance sont normalisés et combinés afin d'obtenir un ensemble restreint de CPO-motifs caractéristiques pour chaque sous-base. La procédure que nous utilisons consiste à sélectionner k motifs, pas-à-pas, en maximisant la valeur suivante, calculée sur le motif à sélectionner P en référence à un ensemble $\mathcal{P}$ de motifs déjà sélectionnés :

$$\mathtt{I}_{\mathtt{PB}}(P, \mathcal{P}) = \frac{\mathtt{NMH}(P, \mathcal{P}) \times \mathtt{Freq}(P) \times \mathtt{NGGR}(P)}{\mathtt{NMH}(P, \mathcal{P}) + \mathtt{Freq}(P) + \mathtt{NGGR}(P)} \times 3 \quad (2)$$

où $\mathtt{NMH}(P, \mathcal{P})$ est la distance de Hamming normalisée minimale entre P et les motifs de l'ensemble $\mathcal{P}$ (distance initialisée à 1 pour $\mathcal{P} = \emptyset$), $\mathtt{Freq}(P)$ est la fréquence de P et $\mathtt{NGGR}(P)$ est son taux de croissance généralisé et normalisé dans la sous-base considérée et vis-à-vis des autres sous-bases. Le facteur 3 est introduit pour des raisons de normalisation.
Cette procédure a une complexité de l'ordre de :

$$\sum_{i=0}^{k-1} i \times (n-i) \times p + (n-i)$$

où n est le nombre de DCPO-motifs extraits de la sous-base B de taille p, et $k > 0$ le nombre de motifs à sélectionner parmi eux. A chaque étape, les i motifs déjà sélectionnés dans $\mathcal{P}$ sont comparés deux à deux aux $n-i$ motifs restants en calculant la distance de Hamming sur la base B.

3 Résultats

Nous ne traitons ici que de l'indice IBGN, qui tient compte de la présence ou de l'absence de certaines espèces macro-invertébrées polluo-sensibles. Parmi les sites de la base de données ne sont donc retenus que ceux sur lesquels l'IBGN a été calculé et pour les périodes où il a été calculé. Le calcul de la note (entre 0, très mauvais état, et 20, très bon état), puis de la classe, est fondé sur l'inventaire de 128 espèces. La méthode est opérée en deux temps : extraction des CPO-motifs discriminants pour chaque sous-base, puis sélection d'un nombre restreint d'entre eux *via* l'indice $\mathtt{I}_{\mathtt{PB}}$. Dans cette expérimentation, le même seuil de fréquence (égal à 10%) est utilisé pour toutes les sous-bases.
La figure 3 donne le nombre (selon une échelle logarithmique) de DCPO-motifs extraits des cinq sous-bases associées aux classes de valeur de l'IBGN. Ces résultats déséquilibrés sont en partie liés aux nombres déséquilibrés de séquences dans les classes, en lien avec l'état observé des cours d'eau : les sous-bases les plus peuplées sont IBGN^V (2405 séquences), lIBGN^J (1282) et IBGN^B (1056). La sous-base IBGN^O est plus faiblement peuplée (556 séquences), tandis que la sous-base IBGN^R est quasi vide (89 séquences). Mais la diversité des motifs extraits des sous-bases s'explique aussi par la diversité des valeurs de para-

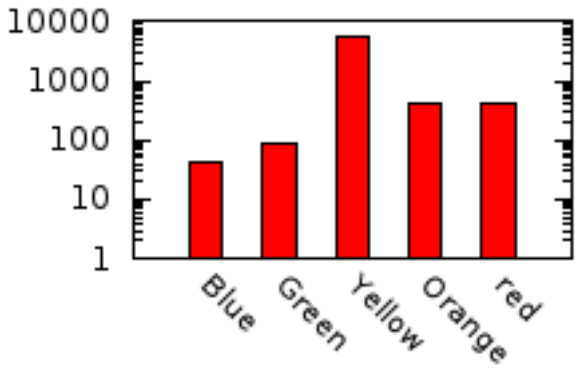

FIGURE 3 – Nombres de CPO-motifs extraits des sous-bases IBGN (classes de valeur *Bleu*, *Vert*, *Jaune*, *Orange* et *Rouge*), avec un support minimum de 1/10

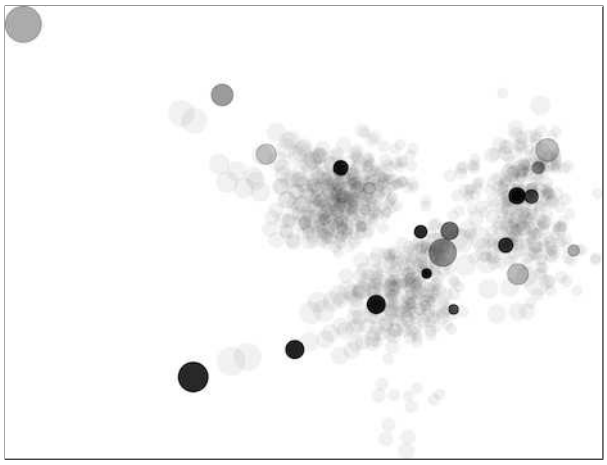

FIGURE 4 – Visualisation dans un plan factoriel des 20 DCPO-motifs sélectionnés parmi les motifs extraits de la sous-base IBGN^R

mètres menant à une même classe d'IBGN. C'est pourquoi il est intéressant de mettre en œuvre la procédure de sélection décrite plus haut, permettant de choisir des DCPO-motifs qui mettent en évidence cette variété.
Nous avons donc sélectionné 20 DCPO-motifs pour chaque sous-base. La figure 4 représente leur distribution vis-à-vis des DCPO-motifs extraits dans la sous-base IBGN^R, dans le plan factoriel construit sur la distance de Hamming entre motifs. Chaque DCPO-motif est représenté par un cercle dont le diamètre représente la fréquence du motif et le niveau de gris la discriminance. On observe sur cette projection que ces 20 DCPO-motifs sont assez bien répartis parmi les motifs extraits de la sous-base, respectant ainsi le critère de non redondance.
Les figures 5 et 6 présentent deux DCPO-motifs extraits des sous-bases associées aux états "Très bon" (IBGN^B) et "Très mauvais" (IBGN^R). Les deux macro-paramètres MINE et MOOX représentent respectivement la minéralisation de l'eau et les matières organiques et oxydables présentes dans l'eau. On remarque ici que les classes des macro-paramètres correspondent exactement aux classes de l'indice biologique. Ce n'est pas forcément le cas pour les autres indices – non présentés dans cet article.

FIGURE 5 – Un DCPO-motif de la sous-base IBGN^B, associant les macro-paramètres "minéralisation de l'eau" et "matières organiques et oxydables"

La figure 7 présente un DCPO-motif caractéristique de la sous-base IBGN^O. On observe ici que différentes séquences mènent à la classe "Mauvais état" de cet indice. Le paramètre TEMP (température) est peu pertinent car peu variable sur les données considérées. En revanche les

FIGURE 6 – Un DCPO-motif de la sous-base IBGN^R (mêmes paramètres mais de classes différentes que dans la figure 5)

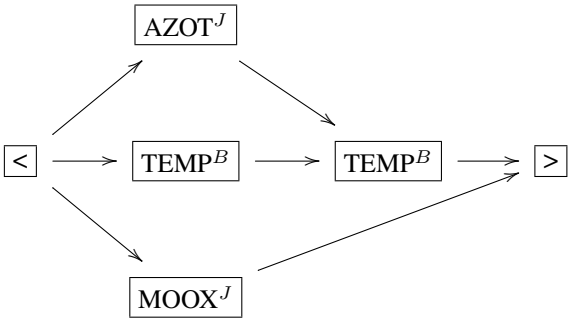

FIGURE 7 – Un DCPO-motif de la sous-base IBGN^O, associant les macro-paramètres "température de l'eau", "matières azotées hors nitrate" et "matières organiques et oxydables"

macro-paramètres AZOT et MOOX – pour des valeurs moyennes (*Jaune*) – ont visiblement un effet dégradant sur l'état mesuré par l'IBGN.

4 Autres travaux

De nombreux travaux ont porté sur l'exploitation de données hydro-écologiques. Classiquement les hydro-écologues utilisent des approches statistiques. Des approches de fouille de données ont également été mises en œuvre, arbres de décision [7], réseaux de neurones [8], ou treillis de Galois [3]. Ces travaux ont généralement pour objectif de mettre en relation des caractéristiques physiques ou physico-chimiques des rivières et les populations de taxons (faune ou flore) qui les habitent. Ainsi, certains auteurs [7] utilisent des arbres de décision pour prédire l'adéquation des habitats (caractérisés par des paramètres physiques et physico-chimiques) d'une rivère en Grèce à certains macro-invertébrés. Le modèle des arbres de régression multiple a été utilisé pour étudier l'impact des conditions physico-chimiques du milieu sur les communautés de diatomées (algues microscopiques) dans un lac macédonien [13]. Avec les mêmes techniques, d'autres auteurs [10] ont cherché à prédire des valeurs de paramètres physico-chimiques à partir de paramètres biologiques (abondance des taxons).
A notre connaissance, aucune étude n'a porté sur l'utilisation d'algorithmes de recherche de motifs dans les séquences, avec l'objectif d'extraire les relations temporelles entre les paramètres physico-chimiques et les indices biologiques, comme nous le proposons ici.

5 Conclusion

Le travail présenté ici a pour but de mettre en évidence les liens temporels entre les valeurs de paramètres physico-chimiques et d'indices biologiques. Nous avons développé pour cela une méthode originale d'extraction de motifs partiellement ordonnés clos. De plus nous avons mis en œuvre une procédure de sélection des motifs s'appuyant sur trois critères combinés, et permettant ainsi d'obtenir des résultats exploitables par l'analyste et représentatifs du jeu de données initial. La méthode a été exploitée avec succès sur un jeu de données important (15 macro-paramètres et environ 5000 séquences pour l'IBGN, mais davantage, de l'ordre de 10.000, quand on considère tous les indices biologiques), ce qui prouve de plus sa pertinence pour la fouille de données massives.
Par la suite, l'étude sera étendue aux données hydro-écologiques collectées sur l'ensemble du territoire français et éventuellement à d'autres indices biologiques ou d'autres paramètres physiques. La méthode devra également être testée sur des jeux de données aux caractéristiques diverses, afin d'en éprouver la stabilité.

Remerciements

Cette recherche est financée par l'Agence Nationale de la Recherche dans le cadre du projet ANR 11 MONU 14 Fresqueau. Nous remercions chaleureusement les hydro-écologues participant au projet, en particulier Corinne Grac (UMR LIVE, ENGEES) et Danielle Levet (Aquascop).

Références

[1] AFNOR. Qualité de l'eau : détermination de l'Indice Biologique Global Normalisé (IBGN). XP T90-350, 2004.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *International Conference on Data Engineering*, ICDE, pages 3–14, 1995.

[3] Aurélie Bertaux, Florence Le Ber, Agnès Braud, and Michèle Trémolières. Identifying Ecological Traits : A Concrete FCA-Based Approach. In *Formal Concept Analysis*, volume 5548, pages 224–236, 2009.

[4] Agnès Braud, Sandra Bringay, Flavie Cernesson, Xavier Dolques, Mickaël Fabrègue, Corinne Grac, Nathalie Lalande, Florence Le Ber, and Maguelonne Teisseire. Une expérience de constitution d'un système d'information multi-sources pour l'étude de la qualité de l'eau. In *Atelier "Systèmes d'Information pour l'environnement", Inforsid 2014, Lyon*, 2014.

[5] Gemma Casas-Garriga. Summarizing Sequential Data with Closed Partial Orders. In *SIAM International Conference on Data Mining, Newport Beach, CA*, SDM PR119, pages 1–12, 2005.

[6] Hong Cheng, Xifeng Yan, Jiawei Han, and Philip S. Yu. Direct Discriminative Pattern Mining for Effective Classification. In *IEEE 24th International Conference on Data Engineering*, ICDE 2008, pages 169–178, 2008.

[7] Eleni Dakou, Tom D'Heygere, Andy P. Dedecker, Peter L.M. Goethals, Maria Lazaridou-Dimitriadou, and Niels Pauw. Decision Tree Models for Prediction of Macroinvertebrate Taxa in the River Axios (Northern Greece). *Aquatic Ecology*, 41 :399–411, 2007.

[8] Andy P. Dedecker, Peter L.M. Goethals, Wim Gabriels, and Niels De Pauw. Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrates in the Zwalm river basin (Flanders, Belgium). *Ecological Modelling*, 174 :161–173, 2004.

[9] Guozhu Dong and Jinyan Li. Efficient Mining of Emerging Patterns : Discovering Trends and Differences. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD, pages 43–52, 1999.

[10] Sašo Džeroski, Damjan Demšar, and Jasna Grbović. Predicting chemical parameters of river water quality from bioindicator data. *Applied Intelligence*, 13(1) :7–17, 2000.

[11] Mickaël Fabrègue, Agnès Braud, Sandra Bringay, Florence Ber, and Maguelonne Teisseire. Orderspan : Mining closed partially ordered patterns. In *Advances in Intelligent Data Analysis XII*, LNCS 8207, pages 186–197. 2013.

[12] Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining : A survey. *ACM Computing Survey*, 38(3), 2006.

[13] Dragi Kocev, Andreja Naumoski, Kosta Mitreski, Svetislav Krstić, and Sašo Džeroski. Learning habitat models for the diatom community in Lake Prespa. *Ecological Modelling*, 221(2) :330–337, 2010.

[14] Jian Pei, Haixun Wang, Jian Liu, Ke Wang, Jianyong Wang, and Philip S. Yu. Discovering Frequent Closed Partial Orders from Strings. *IEEE Transactions on Knowledge and Data Engineering*, 2006.

[15] Miao Wang, Xue-qun Shang, and Zhan-huai Li. Sequential Pattern Mining for Protein Function Prediction. In *Advanced Data Mining and Applications*, volume 5139 of *ADMA*, pages 652–658. 2008.