Exploration de données temporelles avec des treillis relationnels

Cristina Nica, Xavier Dolques, Agnès Braud, Marianne Huchard, Florence Le Ber

Exploration de données temporelles avec des treillis relationnels

Cristina Nica*, Xavier Dolques*, Agnès Braud*
Marianne Huchard**, Florence Le Ber*

*ICube, Université de Strasbourg, CNRS, ENGEES
prenom.nom@engees.unistra.fr, agnes.braud@unistra.fr
http://icube-bfo.unistra.fr
**LIRMM, Université de Montpellier 2, CNRS
huchard@lirmm.fr
https://www.lirmm.fr

Résumé. Cet article présente une méthode d'exploration de données temporelles à l'aide de treillis relationnels. Elle s'applique à un jeu de données composé de séquences de valeurs concernant des paramètres physico-chimiques et biologiques mesurés dans des cours d'eau. L'objectif est d'extraire des sous-séquences fréquentes reliant les deux types de paramètres. Nous montrons sur un petit exemple que l'analyse relationnelle de concepts (ARC) permet de mettre en évidence l'influence dans le temps des paramètres physico-chimiques sur les paramètres biologiques.

1 Introduction

La fouille de données séquentielles (Agrawal et Srikant, 1995) est un domaine de recherche actif qui comporte de nombreuses applications. Par exemple, en étudiant les données acquises auprès de clients d'un magasin, on cherchera à prévoir et favoriser leurs prochains achats ; en étudiant les suivis temporels d'un processus industriel, on cherchera à caractériser les séquences d'évènements qui mènent à un incident. Dans le cas qui nous préoccupe ici, il s'agit d'anticiper et de prévenir l'évolution de l'état des cours d'eau, en examinant les séquences de prélèvements effectués sur les rivières. Plus particulièrement nous essayons de mettre en évidence l'effet dans le temps d'un état physico-chimique de l'eau sur son état biologique (populations animales et végétales) en recherchant des répétitions fréquentes de sous-séquences ayant valeur de règles. Ce travail est mené dans le cadre du projet Fresqueau [1] en interaction avec des hydrobiologistes. Une première approche, s'appuyant sur la recherche de motifs temporels, a été mise en œuvre avec succès par Fabrègue et al. (2014).

Dans cet article, nous explorons une deuxième approche, en considérant ce problème temporel comme une variante d'un problème relationnel. Nous exploitons une technique d'analyse de données relationnelles fondée sur l'analyse de concepts formels (ACF) (Ganter et Wille,

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

1997), l'analyse relationnelle de concepts (ARC), qui est développée depuis une dizaine d'années (Hacene et al., 2013). Cette méthode a été utilisée sur de nombreuses applications mettant en jeux des données relationnelles, en particulier en génie logiciel pour l'analyse d'éléments UML (Arévalo et al., 2006; Dolques et al., 2012) ou pour la détection de défauts (Moha et al., 2008). Cette approche est également utile pour factoriser des classes redondantes, en exploitant les attributs et les relations entre classes (Miralles et al., 2015).

Appliquer l'ARC à l'analyse de données séquentielles requiert une modélisation spécifique que nous allons expliciter dans ce papier à partir d'un exemple simplifié. En revanche nous ne présenterons pas de résultats chiffrés car l'exploitation complète des données pose des difficultés non encore résolues. Dans la suite de l'article nous présenterons ces données, puis les fondements théoriques de la modélisation utilisée ; nous montrerons ensuite sa mise en œuvre et le type de résultats qu'elle permet d'obtenir avant de conclure.

2 Contexte et données

De nombreuses questions de recherche touchant à l'environnement impliquent la prise en compte de données temporelles ou spatiales. Dans le cadre du projet Fresqueau, nous nous intéressons à l'état des cours d'eau, et nous développons des méthodes de fouille de données pour exploiter les données disponibles permettant d'évaluer cet état. Ces données sont de natures et d'origines différentes : elles concernent la qualité de l'eau, l'hydrologie, les stations de mesures, etc. mais également l'occupation du sol au voisinage des cours d'eau (Braud et al., 2014). Les données de qualité de l'eau en particulier sont produites par les agences de l'eau et l'ONEMA (Office National de l'Eau et des Milieux Aquatiques), et 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) (AFNOR, 2004) 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 mesures nationaux. Chaque station d'un réseau de mesures est ainsi caractérisée théoriquement par une note annuelle d'un ou plusieurs indices biologiques, et par des valeurs bimensuelles de paramètres physico-chimiques. Le tableau 1 présente un extrait du jeu de données. À chaque station (colonne 1) sont associées plusieurs dates (sous la forme mois/année, colonne 2) auxquelles sont effectués des prélèvements physico-chimiques (par exemple, NH_4^+, NKJ) ou biologiques (synthétisés sous la forme d'un indice, l'IBGN ici). Ici une seule station est présentée pour des questions de place. Le nombre d'attributs est également limité. On remarque sur ce tableau que les données sont éparses : en particulier, alors que les paramètres physico-chimiques sont mesurés tous les deux à trois mois, l'IBGN est réalisé au mieux une fois l'an, en période d'étiage, donc généralement en été.

Numéro Station	Mois / année	NH_4^+	NKJ	NO_2^-	PO_4^{3-}	Phosphore total	IBGN
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

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

Pour mettre en œuvre la méthode choisie, différents prétraitements sont nécessaires. En particulier, il faut transformer l'information numérique en information qualitative et pour cela nous nous appuyons sur des références du domaine permettant d'agréger les données collectées sur les stations de mesures. De plus, pour tenir compte des connaissances du domaine dans le processus d'analyse, nous ne considérerons par la suite qu'une partie des données. Concrètement, nous éliminons les dates de prélèvements physico-chimiques trop loin dans le temps (au-delà de 4 mois) et avant une date de prélèvement biologique, car nous cherchons à étudier l'effet, limité dans le temps, de l'état physico-chimique du cours d'eau sur la biologie.

Les données ont donc 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" et "Très mauvais" representées par cinq couleurs *Bleu*, *Vert*, *Jaune*, *Orange* et *Rouge*. 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. Les paramètres NH_4^+, NJK et NO_2^- sont eux rassemblés dans le macro-paramètre AZOT (matières azotées hors nitrate).

Le résultat de la discrétisation appliquée au tableau 1 est présenté dans le tableau 2. Une station (*Site 1*) a été ajoutée. Concernant la station *Site 2*, on remarque que le nombre de dates considérées a été réduit de moitié.

3 Analyse de concepts formels

L'analyse de concepts formels (ACF) est une méthode de classification qui s'applique à des jeux de données constitués d'objets décrits par des attributs (Ganter et Wille, 1997). D'un point de vue mathématique, l'ACF permet d'extraire des données un ensemble de concepts munis

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

Site	Date	AZOT	PHOS	IBGN
$Site\ 1$	06/07	Bleu	-	-
	07/07	Vert	Bleu	-
	09/07	-	-	Bleu
	02/08	Bleu	Vert	-
	04/08	Vert	-	-
	05/08	-	-	Jaune
$Site\ 2$	07/04	Orange	Jaune	-
	08/04	Vert	-	-
	09/04	-	-	Orange
	06/05	Bleu	Vert	-
	08/05	-	-	Jaune

TAB. 2 – *Jeu de données discrètes obtenu à partir du tableau 1*

d'une structure hiérarchique. Pour l'illustrer, nous utilisons l'exemple du tableau 2 contenant les données biologiques et physico-chimiques discrétisées.

On considère en entrée du processus ACF un contexte formel qui est un triplet $\mathcal{K} = (O, A, I)$, où O est un ensemble d'objets, A un ensemble d'attributs et I une relation binaire entre O et A, $I \subseteq O \times A$. Le tableau 2 nécessite donc un pré-traitement, qui consiste ici à transformer une relation multi-valuée en une relation binaire. Le contexte $\mathcal{K}_{\texttt{sites}}$ résultant est présenté au tableau 3, avec :

- $O = \{S1\_06/07, S1\_07/07, S1\_09/07, S1\_02/08, S1\_04/08, S1\_05/08, S2\_07/04, S2\_08/04, S2\_09/04, S2\_06/05, S2\_08/05\}$ (site et date sont accolés),
- $A = \{\text{AZOT}^O, \text{AZOT}^J, \text{AZOT}^V, \text{AZOT}^B, \text{PHOS}^J, \text{PHOS}^V, \text{PHOS}^B, \text{IBGN}^O, \text{IBGN}^J, \text{IBGN}^B\}$ (^{O} pour *Orange*, ^{J} pour *Jaune*, ^{V} pour *Vert*, ^{B} pour *Bleu*)
- Les couples de la relation I sont désignés par une croix à la jonction d'une ligne et d'une colonne.

Site-Date	AZOT^O	AZOT^J	AZOT^V	AZOT^B	PHOS^O	PHOS^J	PHOS^V	PHOS^B	IBGN^O	IBGN^J	IBGN^B
$S1\_06/07$				×							
$S1\_07/07$			×					×			
$S1\_09/07$											×
$S1\_02/08$				×			×				
$S1\_04/08$			×								
$S1\_05/08$										×	
$S2\_07/04$	×					×					
$S2\_08/04$			×								
$S2\_09/04$									×		
$S2\_06/05$				×			×				
$S2\_08/05$										×	

TAB. 3 – *Jeu de données binaires obtenu à partir du tableau 2*

On considère maintenant un sous-ensemble d'objets $X \subseteq O$ et un sous-ensemble d'attributs $Y \subseteq A$. On peut alors définir deux opérateurs de Galois, notés $'$, s'appliquant aux sous-ensembles X et Y.

$$X' = \{y \in A | \forall x \in X : (x, y) \in I\} \texttt{ et } Y' = \{x \in O | \forall y \in Y : (x, y) \in I\}$$

Un concept formel est une paire (X, Y) où $X = Y'$ et $Y = X'$. X s'appelle l'extension et Y l'intension du concept. X est l'ensemble maximal d'objets décrits par tous les attributs de Y et Y est l'ensemble maximal d'attributs partagés par tous les objets de X. Par exemple, $\{S1\_06/07\}' = \{\text{AZOT}^B\}$ et $\{\text{AZOT}^B\}' = \{S2\_06/05, S1\_06/07, S1\_02/08\}$: le couple $(\{S2\_06/05, S1\_06/07, S1\_02/08\}, \{\text{AZOT}^B\})$ est un concept formel.

L'ensemble de concepts $\mathcal{C}_K$ construit sur le contexte $\mathcal{K}$ est muni d'un ordre, défini de la façon suivante. Soit $C_1 = (X_1, Y_1)$ et $C_2 = (X_2, Y_2)$ deux concepts formels, $C_1 \leq C_2 \Leftrightarrow X_1 \subseteq X_2 \Leftrightarrow Y_2 \subseteq Y_1$. C_1 est un sous-concept de C_2 et C_2 un sur-concept de C_1. Par exemple le concept $(\{S2\_06/05, S1\_06/07, S1\_02/08\}, \{\text{AZOT}^B\})$ est un sur-concept du concept $(\{S2\_06/05, S1\_02/08\}, \{\text{PHOS}^V, \text{AZOT}^B\})$. L'ensemble $\mathcal{C}_K$ muni de la relation $\leq$ est un treillis de Galois ou treillis de concepts, $\mathcal{L}_K = (\mathcal{C}_K, \leq)$. Le treillis obtenu à partir du contexte $\mathcal{K}_{\texttt{sites}}$ est présenté sur la figure 1.

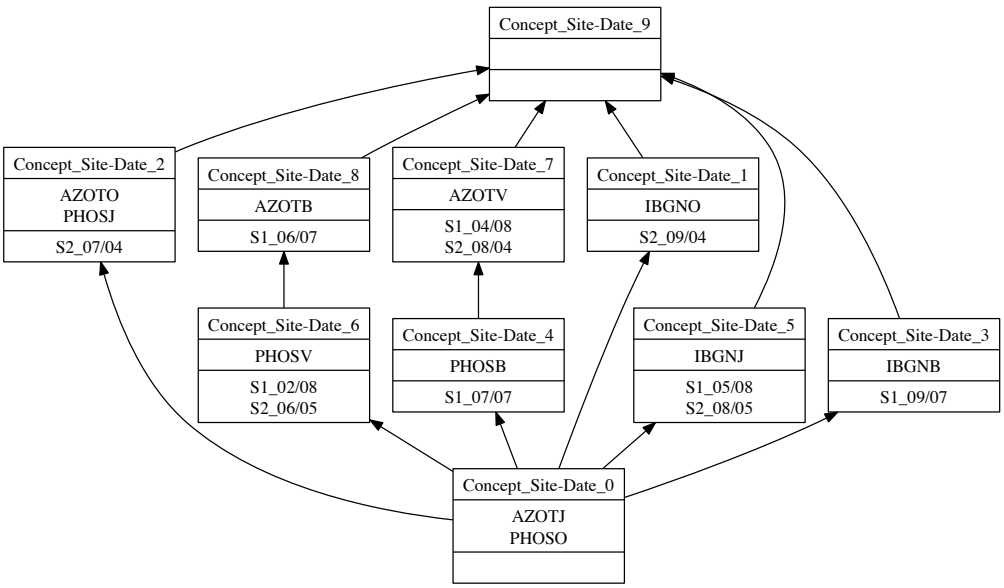

FIG. 1 – *Treillis de concepts obtenu à partir du contexte* $\mathcal{K}_{\texttt{sites}}$

Dans la figuration utilisée, chaque concept est représenté par son extension et son intension simplifiées : seuls les attributs et objets introduits par le concept y sont représentés. Chaque concept hérite des extensions simplifiées de ses sous-concepts et des intensions simplifiées de ses sur-concepts. Par exemple le concept dénommé Concept_Site-Date_6 correspond au concept $(\{S2\_06/05, S1\_02/08\}, \{\text{PHOS}^V, \text{AZOT}^B\})$ décrit ci-dessus ; l'attribut AZOT^B

n'apparaît pas dans l'intension simplifiée, il est hérité du sur-concept Concept_Site-Date_8. Le concept le plus général (en haut dans le treills) a une extension maximale, couvrant tous les objets, tandis que le concept le plus spécifique (en bas dans le treillis) a une intension maximale, couvrant toutes les propriétés. Dans ce treillis particulier, le concept le plus général a une intension vide (aucune propriété n'est partagée par tous les objets) et le concept le plus spécifique a une extension vide (aucun objet ne réunit toutes les propriétés).

4 Analyse relationnelle de données temporelles

Comme montré ci-dessus, l'analyse de concepts formels prend en entrée une unique relation binaire objet-attribut et ne permet donc pas d'exploiter la relation temporelle de précédence inscrite dans le jeu de données traité. C'est pourquoi nous utilisons l'analyse relationnelle de concepts qui permet de prendre en compte des relations inter-objets par application itérative de l'ACF sur un ensemble de contextes formels, qu'on appelle une famille relationnelle de contextes (Hacene et al., 2013). Plus formellement, une famille relationnelle de contextes est composée de n contextes formels objet-attribut $\mathcal{K}_i = (O_i, A_i, I_i)$, $i \in [1..n]$, et de m contextes relationnels objet-objet $\mathcal{R}_j = (O_k, O_l, r_j)$, $j \in [1..m]$ où O_k et O_l sont respectivement des ensembles d'objets de $\mathcal{K}_k$ et $\mathcal{K}_l$ et $r_j \subseteq O_k \times O_l$.

Lors de la première itération du processus ARC, un treillis est généré pour chaque contexte formel $\mathcal{K}_l$. Aux itérations suivantes, les concepts créés à l'étape précédente sont intégrés sous forme d'attributs dans les contextes formels $\mathcal{K}_k$ pour enrichir la description des objets. En effet un concept C_l du treillis $\mathcal{L}_{K_l}$ contient dans son extension un sous-ensemble d'objets de O_l et grâce à une opération d'échelonnage il est possible d'utiliser la relation inter-objets $r_j \subseteq O_k \times O_l$ entre les objets de $\mathcal{K}_k$ et les objets de C_l pour créer une relation objet-concept. Dans la suite nous utiliserons l'opérateur *existentiel* qui crée une relation $\exists r_j$ entre un objet $o \in O_k$ et un concept C_l dès que $r_j(o)$ a une intersection non vide avec l'extension de C_l. L'attribut $\exists r_j(C_l)$ est alors ajouté au contexte $\mathcal{K}_k$.

Au cours du processus, chaque contexte formel ainsi augmenté permet de générer un nouveau treillis où les attributs ajoutés peuvent faire émerger de nouveaux concepts. Lorsqu'aucun nouveau concept n'apparaît lors d'une nouvelle itération, le processus de l'ARC a atteint un point fixe et s'arrête.

4.1 Modélisation des données

Comme nous l'avons dit ci-dessus, le processus ARC prend en entrée une famille relationnelle de contextes qui contient un ensemble de contextes formels et un ensemble de contextes relationnels. La création d'une telle famille relationnelle de contextes implique d'évaluer la connaissance à obtenir et le type de la donnée initiale (par exemple temporelle, spatiale). Dans l'exemple traité, l'objectif est d'analyser l'influence des paramètres physico-chimiques sur les paramètres biologiques. La phase d'analyse s'appuie sur les aspects temporels, c'est-à-dire sur l'évolution de la qualité des paramètres prélevés par station durant une période de temps spécifique. Deux types de relations sont considérés, les relations temporelles (un prélèvement précède un autre prélèvement) et les relations de valeur (un prélèvement a pour valeur la classe *Jaune* ou *Verte*, etc.). Quatre contextes formels doivent donc être créés : le

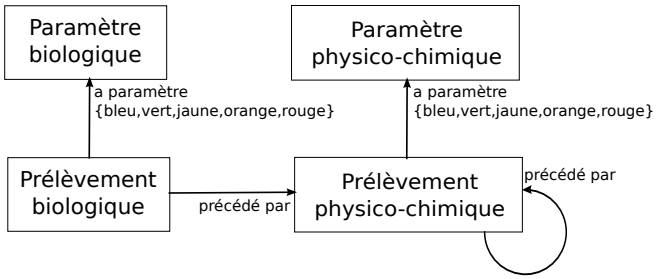

FIG. 2 – *Schéma relationnel illustrant la modélisation de la famille relationnelle de contextes*

contexte des paramètres biologiques (réduit ici à l'IBGN), le contexte des paramètres physico-chimiques (AZOT, PHOS), le contexte des prélèvements biologiques et le contexte des prélèvements physico-chimiques. La figure 2 illustre la mise en relation de ces différents contextes par les relations temporelles (entre contextes des prélèvements) et les relations de valeur (entre contextes des prélèvements et des paramètres).

Reconsidérons le tableau 2. Les contextes objet-attribut sont construits directement à partir des en-têtes ligne et colonne de ce tableau : le tableau 4 montre le contexte des paramètres biologiques, $\mathcal{K}_{paramBio}$ (à gauche), et le contexte des prélèvements biologiques, $\mathcal{K}_{prelBio}$ (au milieu). Le contexte des paramètres a pour particularité d'avoir un ensemble d'objets et un ensemble d'attributs identiques ($\{IBGN\}$ dans l'exemple). Le contexte des prélèvements a un ensemble d'attributs vide.

La construction des contextes relationnels exploite quant à elle les relations sous-jacentes au tableau 2. Celui-ci contient des données multi-valuées, c'est-à-dire qu'un paramètre y est décrit par différentes valeurs de qualité. À chaque valeur correspond un contexte relationnel entre les prélèvements biologiques et les paramètres biologiques ; de même pour les prélèvements physico-chimiques et les paramètres physico-chimiques. De manière concise, il y a sept relations binaires : $R_{aParamYX} = \mathcal{K}_{prelX} \times \mathcal{K}_{paramX}$ où $Y \in \{Vert, Bleu, Orange, Jaune\}$ si $X \in \{bio\}$ ou $Y \in \{Bleu, Jaune, Orange\}$ si $X \in \{phc\}$. Le tableau 4 (à droite) illustre un des contextes relationnels mentionnés précédemment. Enfin, les dates associées aux prélè-

TAB. 4 – *De gauche à droite : contexte des paramètres biologiques, contexte des prélèvements biologiques et contexte relationnel des prélèvements biologiques ayant la valeur Jaune*

$\mathcal{K}_{paramBio}$	IBGN
IBGN	×

$\mathcal{K}_{prelBio}$
S1_09/07
S1_05/08
S2_09/04
S2_08/05

$\mathcal{R}_{aParamJBio}$	IBGN
S1_09/07	
S1_05/08	×
S2_09/04	
S2_08/05	×

vements donnent une information sur les relations temporelles entre les contextes formels de prélèvements. La relation $\mathcal{R}_{bioPrecedeParPhC} = \mathcal{K}_{prelBio} \times \mathcal{K}_{prelPhC}$, exprime le fait qu'un paramètre physico-chimique a été prélevé avant un paramètre biologique sur une même station. La relation $\mathcal{R}_{phcPrecedeParPhC} = \mathcal{K}_{prelPhC} \times \mathcal{K}_{prelPhC}$ est un contexte relationnel cy-

clique. Elle indique qu'un paramètre physico-chimique a été prélevé avant un autre paramètre physico-chimique sur une même station.

Le processus ARC est orienté de la façon suivante. L'ACF est appliquée une seule fois sur les contextes de paramètres $\mathcal{K}_{paramBio}$ et $\mathcal{K}_{paramPhC}$ dont les ensembles objets et attributs sont identiques. Les contextes des prélèvements $\mathcal{K}_{prelBio}$ et $\mathcal{K}_{prelPhC}$, qui ne contiennent que des objets au départ, évoluent au cours du processus par ajout d'attributs relationnels, ce qui permet de mettre en évidence des répétitions, sur différentes stations, d'enchaînements de prélèvements et de leurs valeurs.

4.2 Mise en œuvre de l'ARC

Dans cette partie nous exploitons la famille de contextes relationnels décrite ci-dessus pour contruire un ensemble de treillis permettant de relier les contextes.

Tout d'abord, la procédure itérative de l'ARC commence par l'application d'un algorithme standard pour créer des hiérarchies de concepts à partir des quatre contextes formels $\mathcal{K}_{paramBio}$, $\mathcal{K}_{paramPhC}$, $\mathcal{K}_{prelBio}$ et $\mathcal{K}_{prelPhC}$. La figure 3 présente les résultats de cette première étape, résultats qui servent d'entrée à l'étape suivante.

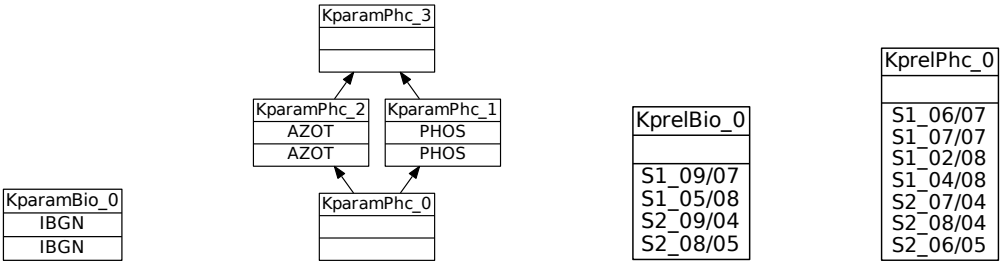

FIG. 3 – *De gauche à droite : treillis issus des contextes formels $\mathcal{K}_{paramBio}$, $\mathcal{K}_{paramPhC}$, $\mathcal{K}_{prelBio}$ et $\mathcal{K}_{prelPhC}$ lors de la première étape du processus ARC*

Dans un deuxième temps, les treillis de la figure 3 sont utilisés pour générer une nouvelle famille de contextes en utilisant l'opérateur d'échelonnage existentiel sur les relations $\mathcal{R}_{bioPrecedeParPhC}$, $\mathcal{R}_{phcPrecedeParPhC}$ et les contextes relationnels qui donnent la qualité des paramètres physico-chimiques et biologiques pour chaque échantillonnage. Le tableau 5 décrit le contexte augmenté $\mathcal{K}_{prelBio}+$ obtenu à cette deuxième étape. Brièvement, le contexte $\mathcal{K}_{prelBio}$ est étendu avec deux types d'attributs relationnels : les attributs de la forme générale $\exists\mathcal{R}_{aParamXBio(KparamBio\_0)}$, où $X \in \{Orange, Jaune, Bleu\}$ représentent le fait qu'un prélèvement biologique appartient à la classe de qualité X de l'indice IBGN ; l'attribut $\exists\mathcal{R}_{bioPrecedeParPhC(KprelPhC\_0)}$ dénote quant à lui qu'il y a au moins un prélèvement physico-chimique du concept $KprelPhC\_0$ précédant le prélèvement biologique considéré.

Le processus se poursuit de la même façon jusqu'à un point fixe obtenu ici après deux autres étapes. La figure 4 présente les treillis obtenus à l'issue du processus complet, durant lequel les deux treillis des paramètres biologiques et physico-chimiques ne changent pas.

Dans le treillis $\mathcal{L}_{\mathcal{K}prelBio}$, le concept $KprelBio\_3$ révèle que la classe de qualité *Bleu* de l'indice IBGN est influencée, durant 4 mois, par la classe de qualité *Bleu* du macro-paramètre

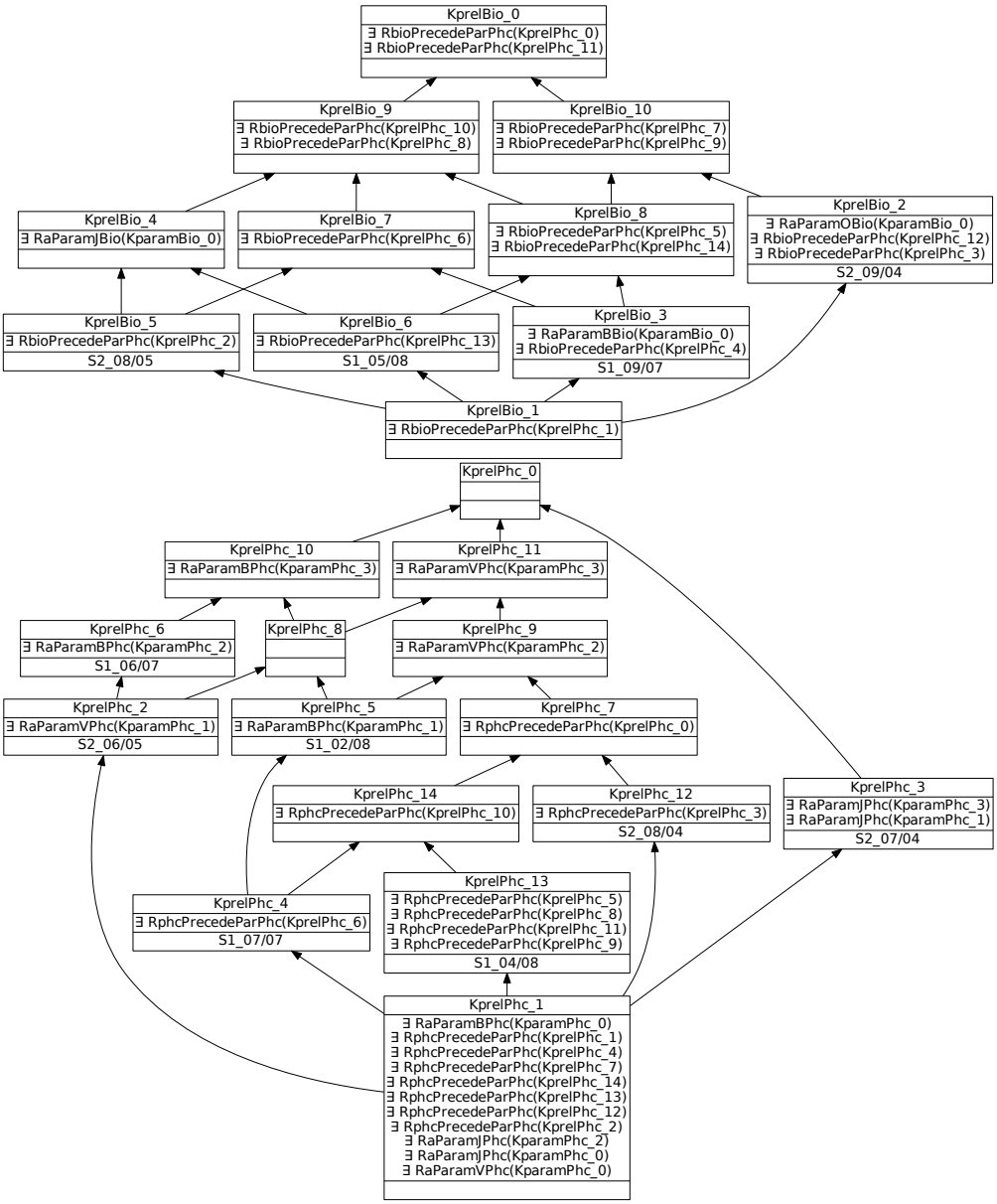

FIG. 4 – *Les deux treillis $\mathcal{L}_{\mathcal{K}prelBio}$ (en haut) et $\mathcal{L}_{\mathcal{K}prelPhC}$ (en bas) obtenus à la dernière étape du processus ARC (comparer avec leur forme initiale en figure 3)*

TAB. 5 – *Contexte augmenté $\mathcal{K}_{prelBio}+$ obtenu à la deuxième étape du processus ARC*

KprelBio	∃ **RaParamOBio(KparamBio_0)**	∃ **RaParamBBio(KparamBio_0)**	∃ **RbioPrecedeParPhc(KprelPhc_0)**	∃ **RaParamJBio(KparamBio_0)**
S1_09/07		×	×	
S1_05/08			×	×
S2_09/04	×		×	
S2_08/05			×	×

PHOS et par la classe de qualité *Vert* du macro-paramètre AZOT. De plus ces paramètres ont été affectés par la classe de qualité *Bleu* du macro-paramètre AZOT. La figure 5 montre l'enchaînement de concepts qui permet d'expliquer ce concept $KprelBio\_3$.

Parallèlement, le treillis $\mathcal{L}_{\mathcal{K}_{prelPhC}}$ contient le concept $KprelPhC\_12$ qui regroupe les prélèvements physico-chimiques décrits par la classe de qualité *Vert* du macro-paramètre AZOT. Cette valeur est influencée par la classe de qualité *Jaune* du même paramètre et du macro-paramètre PHOS.

Pour résumer, on voit dans ce petit exemple que l'ARC peut être utilisée pour révéler des relations temporelles entre paramètres physico-chimiques, qui ont une influence à court terme sur les valeurs des indices biologiques. Les résultats obtenus ici n'ont aucune valeur statistique compte tenu de la petitesse du jeu de données exemple utilisé.

5 Discussion et conclusion

De nombreux travaux ont porté sur l'exploitation de données hydroécologiques, que ce soit avec des approches statistiques classiques ou plus récemment, des méthodes d'apprentissage automatique. Les treillis ont été utilisés par (Bertaux et al., 2009), avec pour objectif d'aider l'expert à la constitution de groupes de taxons (ici des plantes aquatiques) partageant des caractéristiques communes. Les travaux exploitant les techniques d'apprentissage 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, Dakou et al. (2007) utilisent des arbres de décision pour prédire l'adéquation des habitats à certains macro-invertébrés, tandis que Kocev et al. (2010) utilisent les arbres de régression multiple pour étudier l'impact des conditions physico-chimiques du milieu sur des communautés d'algues

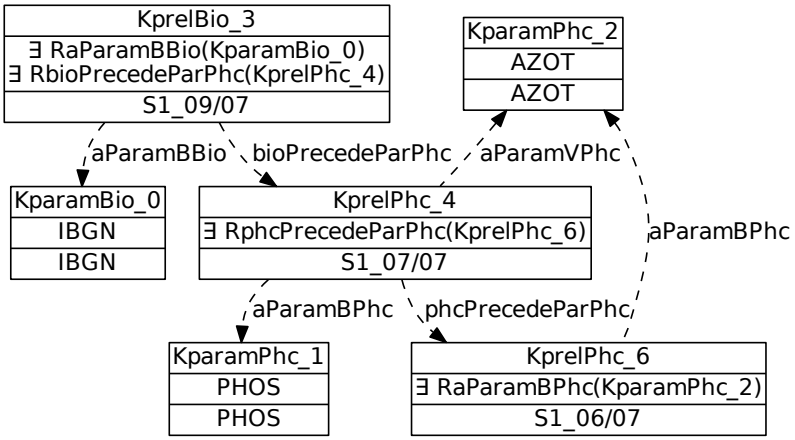

FIG. 5 – *Analyse du concept $KprelBio\_3$: liens entre concepts à travers les différents treillis*

microscopiques. Avec les mêmes techniques, Džeroski et al. (2000) ont cherché à prédire des valeurs de paramètres physico-chimiques à partir de paramètres biologiques (abondance des taxons). En revanche, à notre connaissance, les aspects temporels ne sont pas pris en compte comme nous l'avons fait ici et comme l'ont fait auparavant Fabrègue et al. (2014).

Par rapport à ce dernier travail, qui s'appuie sur la recherche de motifs partiellement ordonnés dans les séquences de données hydroécologiques, nous avons ici mis en œuvre une approche originale pour l'exploration de données temporelles fondée sur l'analyse relationnelle de concepts. Dans un délai proche, nous explorerons le même jeu de données que celui utilisé par Fabrègue et al. (2014). Sa taille importante conduira à des problèmes d'échelle, que nous pourrons résoudre en le segmentant selon les points d'intérêt des utilisateurs (par exemple selon les classes de valeurs des indices biologiques) ou en filtrant les concepts construits selon un seuil appliqué sur la taille de leur extension, ce d'autant que nous sommes à la recherche de sous-séquences fréquentes. Nous pourrons alors mener une comparaison complète avec les résultats obtenus par recherche de motifs.

Remerciements

Cette recherche est financée par l'Agence Nationale de la Recherche dans le cadre du projet ANR 11 MONU 14 Fresqueau.

Références

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

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In *International Conference on Data Engineering*, ICDE, pp. 3–14.

Arévalo, G., J.-R. Falleri, M. Huchard, et C. Nebut (2006). Building Abstractions in Class Models : Formal Concept Analysis in a Model-Driven Approach. In *MoDELS 2006*, pp. 513–527.

Bertaux, A., F. Le Ber, A. Braud, et M. Trémolières (2009). Identifying Ecological Traits : A Concrete FCA-Based Approach. In *Formal Concept Analysis*, LNAI 5548, pp. 224–236.

Braud, A., S. Bringay, F. Cernesson, X. Dolques, M. Fabrègue, C. Grac, N. Lalande, F. Le Ber, et M. Teisseire (2014). 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*.

Dakou, E., T. D'Heygere, A. P. Dedecker, P. L. Goethals, M. Lazaridou-Dimitriadou, et N. Pauw (2007). Decision Tree Models for Prediction of Macroinvertebrate Taxa in the River Axios (Northern Greece). *Aquatic Ecology 41*, 399–411.

Dolques, X., M. Huchard, C. Nebut, et P. Reitz (2012). Fixing Generalization Defects in UML Use Case Diagrams. *Fundam. Inform. 115*(4), 327–356.

Džeroski, S., D. Demšar, et J. Grbović (2000). Predicting chemical parameters of river water quality from bioindicator data. *Applied Intelligence 13*(1), 7–17.

Fabrègue, M., A. Braud, S. Bringay, C. Grac, F. Le Ber, D. Levet, et M. Teisseire (2014). Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. *Ecological Informatics 24*, 210–221.

Ganter, B. et R. Wille (1997). *Formal Concept Analysis : Mathematical Foundations* (1st ed.). Secaucus, NJ, USA : Springer-Verlag New York, Inc.

Hacene, M. R., M. Huchard, A. Napoli, et P. Valtchev (2013). Relational concept analysis : mining concept lattices from multi-relational data. *Ann. Math. Artif. Intell. 67*(1), 81–108.

Kocev, D., A. Naumoski, K. Mitreski, S. Krstić, et S. Džeroski (2010). Learning habitat models for the diatom community in Lake Prespa. *Ecological Modelling 221*(2), 330–337.

Miralles, A., M. Huchard, X. Dolques, F. Le Ber, T. Libourel, C. Nebut, et A. Osman-Guédi (2015). Méthode de factorisation progressive pour accroître l'abstraction d'un modèle de classes. *Revue d'Ingéniérie des Systèmes d'Information*. À paraître.

Moha, N., A. R. Hacene, P. Valtchev, et Y.-G. Guéhéneuc (2008). Refactorings of Design Defects Using Relational Concept Analysis. In *ICFCA 2008*, pp. 289–304.

Summary

This article describes a temporal data mining method based on relational lattices. This method is applied on a sequence dataset, dealing with physico-chemical and biological parameters sampled in watercourses. Our aim is to reveal frequent sub-sequences linking the two parameter types. We use a small example to show that relational concept analysis (RCA) is able to highlight temporal impacts of physico-chemical parameters on biological parameters.