Recherche de séquences spatio-temporelles peu contredites dans des données hydrologiques

Hugo Alatrista Salas[*,♯], Jérôme Azé[**], Sandra Bringay[***,****], Flavie Cernesson[*], Frédéric Flouvat[♯], Nazha Selmaoui-Folcher[♯], Maguelonne Teisseire[*,***]

[*]UMR TETIS, 500 rue Jean-François Breton, F-34093 Montpellier
{prénom.nom}@teledetection.fr
[**]LRI, Équipe Bioinformatique, Projet INRIA AMIB,
Université Paris-Sud 11, 91405 Orsay Cedex
aze@lri.fr
[***]LIRMM, UMR 5506, 161 rue Ada 34392 Montpellier
bringay@lirmm.fr
[****]Département MIAp, UM3 Université Paul-Valéry,
Route de Mende, 34000 Montpellier
[♯]PPME - Univ. de la Nouvelle-Calédonie, F-98851 Nouméa, Nouvelle-Calédonie
{frederic.flouvat,nazha.selmaoui}@univ-nc.nc

Résumé. Dans cet article, nous présentons un projet de découverte de connaissances dans des données hydrologiques. Pour cela, nous appliquons un algorithme d'extraction de motifs séquentiels sur les données relevées au niveau de stations réparties le long de plusieurs rivières. Les données sont pré-traitées afin de considérer différentes proximités spatiales et l'analyse du nombre de motifs obtenus souligne l'influence des relations ainsi définies. Nous proposons et détaillons une mesure objective d'évaluation, appelée la mesure de moindre contradiction temporelle, afin d'aider l'expert dans la découverte de nouveautés. Ces éléments posent les premières bases de travaux plus ambitieux permettant de proposer des indicateurs spatialisés pour l'aide à l'interprétation des données de suivi écologique des cours d'eau et des données de pression.

1 Introduction

Le réseau hydrographique, structurant paysages et écosystèmes, compte plus de $500\,000$ kilomètres pour la France métropolitaine. Le réseau hydrographique est un milieu fragile soumis à la présence de nombreuses activités économiques et des usages qui ont modifié, au cours du temps, son intégrité physique et altéré la qualité physico-chimique et biologique de l'eau. Or, les nouvelles réglementations européenne et nationale affichent explicitement la préservation et la restauration des milieux dont les cours d'eau et la demande sociale inscrit dans sa qualité de vie, la qualité des milieux qui l'environne. Si des dispositifs de suivi de la qualité de l'eau ont été mis en place depuis plusieurs décennies, il s'agit maintenant de construire des indicateurs permettant de rendre compte de l'influence des usages et des

mesures de restauration sur la qualité de l'eau. Pour arriver à la construction de tels outils, il faut prendre en compte différents types de données : (1) les données hydrologiques, ici, liées à la qualité de l'eau (2) les données relatives aux stations de mesure (localisation, réseau d'appartenance ...) (3) le réseau hydrographique, ses caractéristiques physiques et les espaces qui lui sont associés : bassin versant, masse d'eau ... (4) les données relatives aux activités humaines ; et finalement (5) les variables de forçage ou de contexte telles que les données climatiques, ou les données rendant compte d'homogénéité hydroécologique (comme les hydroécorégions). Les données ainsi produites constituent une source d'information volumineuse qu'il est difficile d'analyser de façon globale. L'hétérogénéité et la quantité des données manipulées nécessite la mise en œuvre d'approches spécifiques. Ce besoin est d'autant plus important que les données évoluent suivant un axe temporel.

Dans cet article, nous souhaitons explorer les données de qualité d'eau (ici des indicateurs biologiques) en prenant en compte les informations spatiales (e.g. localisation de stations le long des rivières, leur proximité...) tout en se préoccupant de l'évolution temporelle des données considérées. L'approche que nous adoptons est composée de quatre phases (1) la préparation des données, (2) l'extraction de motifs prenant en compte l'aspect temporel des données (3) la sélection des motifs extraits selon une mesure d'intérêt temporel et enfin (4) la restitution à partir d'un regroupement de ceux-ci en groupes similaires. Ce processus général est illustré sur la figure 2. De façon plus détaillée, une série de pré-traitements va permettre de traduire la prise en compte de différentes proximités spatiales (e.g. rapprochement des stations selon leur distance, selon leur appartenance à une zone...). Pour retracer la prise en compte de l'historique des données, nous appliquons une méthode de recherche de motifs séquentiels définis par (Agrawal et Srikant, 1995). Ces motifs correspondent à des séquences fréquentes d'états mesurés par les stations. Nous définissons ensuite une mesure de pertinence associée à *la moindre contradiction temporelle* du motif considéré par rapport à l'ensemble des motifs extraits. Nous recherchons ainsi les motifs qui sont associés aux mêmes états mais dans des ordres chronologiques différents. À notre connaissance, il s'agit de la première proposition d'une mesure d'intérêt se préoccupant de l'ordre temporel d'apparition des événements. Les résultats sont restitués selon deux mesures objectives, le support et la moindre contradiction temporelle afin d'assister les experts dans leur découverte de nouvelles connaissances. De manière à offrir un outil d'analyse et d'exploration des données hydrologiques et spatiales aux experts, nous proposons également un regroupement des séquences basé sur la mesure de similarité définie par (Saneifar et al., 2008). Ce regroupement permet de considérer ensemble les séquences les plus similaires et représente ainsi une aide à la découverte.

Dans la suite de cet article, nous présentons, section 2, les données manipulées. Nous décrivons ensuite la méthode de fouille de données adoptée dans la section 3. Nous détaillons les pré-traitements réalisés pour prendre en compte la spatialisation dans la section 4.1. La mesure de moindre contradiction temporelle utilisée pour filtrer les motifs est définie section 4.3 ainsi que l'algorithme associé. La méthode de regroupement de séquences utilisée pour la restitution des résultats est présentée, section 4.4. Les expérimentations réalisées sont retracées, section 5. Nous dressons ensuite le bilan de ces propositions en soulignant les perspectives à court et moyen terme.

2 Les données

La base de données considérée est constituée de relevés d'indicateurs biologiques mesurés sur la Saône et ses affluents.

La figure 1 décrit le positionnement géographique des cours d'eau et des stations de relevés dans le bassin versant considéré. Les données à notre disposition se présentent sous deux formes : les données statiques qui caractérisent les stations de prélèvement des échantillons et les données dynamiques qui sont issues des prélèvements et permettent après analyse de calculer les indices biologiques, ou qui sont directement cet indice biologique. L'ensemble des données référantes aux rivières sont décrites dans le tableau 1.

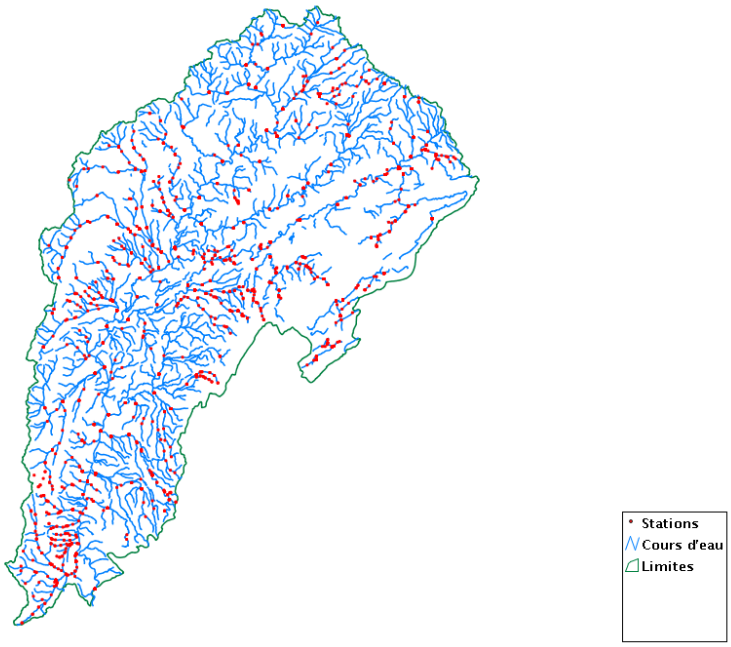

FIG. 1 – *Le bassin versant de la Saône*

Les données statiques sont associées aux stations de relevés positionnées sur les cours d'eau. Chaque station (codstace) est décrite par :

- Les coordonnées Lambert (x,y) pour identifier la position spatiale de chaque station de prélèvement identifiée par *codstace*. Le système de projection Lambert 93 est ici adopté pour effectuer le géo-référencement ;
- Un point kilométrique - grandeur utilisée pour localiser un point le long d'un cours d'eau et qui est calculée en mesurant, en kilomètres, la portion du cours comprise entre le point localisé et un point servant d'origine (la confluence) ;
- Une hydro-écorégion - unité spatiale homogène du point de vue de la géologie, du relief et du climat. C'est l'un des principaux critères utilisés dans la typologie et la délimitation

des masses d'eau de surface. La France métropolitaine est décomposée en 22 hydroécorégions et 7 hydroécorégions sont présentes sur la zone étudiée (hydroecor) ;
- Le codmasseau pour codifier les masses d'eau correspondant ici à des eaux superficielles telles qu'une rivière ou un canal, une partie de rivière, de fleuve ou de canal. Pour la Saône, il existe un total de 572 masses d'eau "type cours d'eau". En revanche, nous ne traitons pas les masses d'eau tels que les lacs et plans d'eau ;
- La taille des masses d'eau (Très Petit, Petit, ... , Très Grand) définie par la position de la masse d'eau dans le réseau hydrographique ;
- Un contexte piscicole - unité spatiale dans laquelle une population de poissons fonctionne de façon autonome.

Les données dynamiques correspondent aux relevés effectués par les stations (*cf.* tableau1). La fréquence de ces relevés, ainsi que les informations contrôlées, varient en fonction du temps et des stations. Certaines stations possèdent des relevés récurrents alors que d'autres stations ne présentent qu'un seul relevé effectué, par exemple, dans le cadre d'une étude ponctuelle.

Les principales informations associées aux relevés sont les suivantes :
- La date du relevé ;
- L'IBGN : Indice Biologique Global Normalisé (calcul standardisé basé sur la détermination des macroinvertébrés d'eau douce) ;
- L'IBD : Indice Biologique Diatomée (calcul standardisé de diagnostic des pollutions trophiques).

Les indicateurs IBGN et IBD sont normalisés en fonction de la masse d'eau étudiée et de l'hydro-écorégion. Trois notes sont alors obtenues et comparables entre les différentes stations : une note pour l'IGBN, une note pour l'IBD et une note correspondant à la fusion normalisée des deux notes précédentes. Cette dernière information permet d'estimer l'état du cours d'eau au niveau du point de relevé.

codstace	codmasseau	x	y	hydroecor	rdate	ibgn	ibd
6000890	FRDR696	863500	2332140	10	2008-09-23	-100	12
6000890	FRDR696	863500	2332140	10	2001-07-10	17	-100
6000950	FRDR694	893478	2346387	4	2008-08-28	17	13
6000980	FRDR697	866447	2341582	10	2008-08-27	15	12
6001250	FRDR691	864725	2323175	10	2003-08-20	-100	12.5
6003550	FRDR680	877007	2300933	10	2008-07-31	-100	14
6456610	FRDR631	946436	2295348	18	2008-07-19	-100	12.3
...	...	...	...	...	...	...	...

TAB. 1 – *Données des relevés*

3 Extraction de motifs séquentiels : panorama et définitions

Dans cette section, nous introduisons les définitions relatives à l'extraction de motifs séquentiels (Agrawal et Srikant, 1995).

Le problème de la recherche de motifs séquentiels a été introduit par (Agrawal et Srikant, 1995) dans le contexte du panier de la ménagère et appliqué avec succès dans de nombreux domaines comme la biologie (Wang et al., 2004; Salle et al., 2009), la fouille d'usage du Web (Pei et al., 2000; Masseglia et al., 2008), la détection d'anomalie (Rabatel et al., 2010), la

fouille de flux de données (Marascu et Masseglia, 2006) ou la description des comportements au sein d'un groupe (Perera et al., 2009). Des approches plus récentes (Julea. et al., 2008) utilisent les motifs séquentiels pour décrire les évolutions temporelles des pixels au sein des séries d'images satellites.

Considérons la base DB, illustrée tableau 2, regroupant l'ensemble des achats réalisés par des clients. Chaque n-uplet T correspond à une transaction financière et consiste en un triplet *(id-client, id-date, itemset)* : l'identifiant du client, la date de l'achat ainsi que l'ensemble des produits (items) achetés.

Soit $I = \{i_1, i_2, \ldots, i_m\}$ l'ensemble des $items$ (produits). Un *itemset* est un ensemble d'items non vide noté $(i_1, i_2, \ldots, i_k)$ où i_j est un *item* (il s'agit d'une représentation non ordonnée). Une *séquence* S est une liste ordonnée, non vide, d'itemsets notée $< s_1 s_2 \ldots s_p >$ où s_j est un itemset. Une *n-séquence* est une séquence composée de n items. Par exemple, considérons les achats des produits A, B, C, D et E réalisés par le client Dupont selon la séquence $S =< (A, E)(B, C)(D)(E) >$ indiquée dans le tableau 2. Ceci signifie qu'hormis les achats des produits A et E puis B et C qui ont été réalisés ensemble, i.e. lors de la même transaction, les autres items de la séquence ont été achetés plus tard et séparément. Dans notre exemple, S est une 6-séquence.
Une séquence $< s_1 s_2 \ldots s_p >$ est une sous-séquence d'une autre séquence $< s'_1 s'_2 \ldots s'_m >$ s'il existe des entiers $i_1 < i_2 < \ldots < i_j < \ldots < i_p$ tels que $s_1 \subseteq s'_{i_1}, s_2 \subseteq s'_{i_2}, \ldots, s_p \subseteq s'_{i_p}$. Par exemple, la séquence S' = $< (B)(E) >$ est une sous-séquence de S car $(B) \subseteq (B, C)$ et $(E) \subseteq (E)$. Toutefois, $< (B)(C) >$ n'est pas une sous-séquence de S car les deux itemsets de S' ne sont pas inclus dans deux itemsets de S.

Tous les achats d'un même client sont regroupés et triés par date. Ils constituent la séquence de données du client. Un client *supporte* une séquence S si S est incluse dans la séquence de données de ce client (S est une sous-séquence de la séquence de données). Le *support* d'une séquence S est alors calculé comme étant le pourcentage des clients qui supportent S. Soit $minSupp$ le support minimum fixé par l'utilisateur, une séquence qui vérifie le support minimum (i.e. dont le support est supérieur à $minSupp$) est une *séquence fréquente*.

Client	Date	Items
Dupont	04/01/12	TV(A), Chocolat (E)
Martin	04/02/28	Chocolat (E)
Dupont	04/03/02	Lecteur DVD (B) , Caméscope (C)
Dupont	04/03/12	Imprimante (D)
Dupont	04/04/26	Chocolat (E)

TAB. 2 – *Exemple d'une base d'achats*

Le problème de la recherche de motifs séquentiels dans une base de données consiste à trouver les séquences maximales dont le support est supérieur au support minimum spécifié, noté $minSupp$. Chacune de ces séquences est un motif séquentiel ou plus communément une séquence fréquente.

Transposer la problématique d'extraction de motifs séquentiels aux données hydrologiques revient tout d'abord à discrétiser les valeurs numériques et à définir quel est le critère de comptage pour la fréquence. L'approche la plus naïve serait de choisi les stations de relevés pour un tel critère. Dans ce cas, un exemple de motif séquentiel pourrait être : $< (ibgn\_etat\_TBE)(ibgn\_0\text{-}10, gr\_indic\_0.4) >, 30\%$ qui signifie que les mesures indi-

quées ont été relevées pour 30% des stations. Une limite à cette approche est un nombre de séquences fréquentes très faible. Ceci est dû essentiellement au caractère hétérogène des relevés selon les stations. Nous avons donc proposer d'intégrer un autre critère de comptage à partir du critère spatial permettant de regrouper plusieurs stations. Ceci permet d'obtenir plus de séquences fréquentes. Ces points seront développés dans la section 4.1.

Il existe de nombreuses propositions de fouille de données dîtes spatio-temporelles c'est-à-dire intégrant à la fois les aspects temporels et spatiaux (Cao et al., 2005; Celik et al., 2006; Mabit et al., 2011). Néanmoins, à notre connaissance, l'étude de la litérature ne fait état d'aucun travaux sur l'application de techniques de recherche de séquences sur des séries temporelles en y incluant des opérateurs spatiaux adaptés au contexte de données hydrologiques.

4 Le processus de découverte de connaissances appliqué aux données hydrologiques

La figure 2 montre le schéma du processus de découverte des connaissances appliqué aux données du bassin versant de la Saône. Les étapes de ce processus sont décrites dans les sous-sections suivantes.

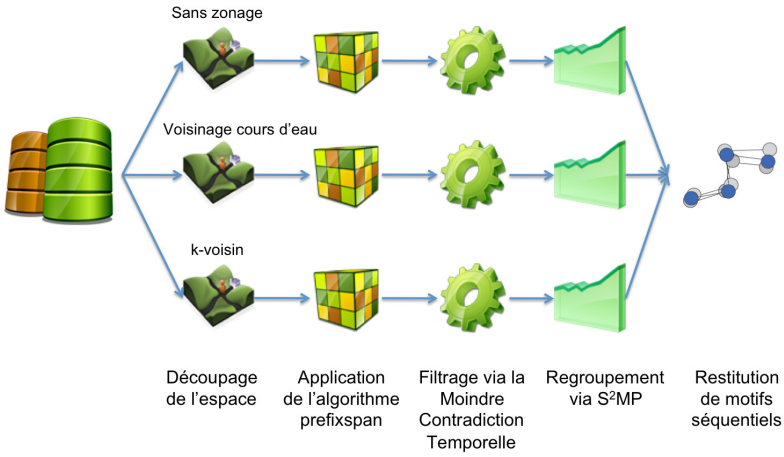

FIG. 2 – *Processus de découverte de connaissances appliqué aux données hydrographiques*

4.1 Le pré-traitement des données : intégration de la dimension spatiale

Cette section détaille comment la dimension spatiale a été intégrée dans notre approche, permettant ainsi la découverte de séquences spatio-temporelles adaptées à l'étude des cours d'eau.

Les données spatiales nous permettent de déterminer des zones géographiques pertinentes afin de gérer (a) la proximité à partir de la localisation des stations exprimée par leurs coordonnées Lambert (système de coordonnées géoréférencées) ; (b) les aspects écoulement alliant

la proximité liée au cours d'eau, le sens de circulation ainsi que les connexions entre les cours d'eau.

Nous avons pré-traité les données afin de découper l'espace en zones. Les séquences de données sont ensuite obtenues en regroupant les données d'une même zone et en les triant par date. Ainsi, nous pouvons utiliser un algorithme classique d'extraction de séquences intéressantes, comme expliqué dans la section 4.2 pour découvrir des séquences temporelles.

Dans cet article, deux découpages de l'espace sont proposés :

1. Un voisinage restreint au cours d'eau : pour un cours d'eau donné, deux stations X et Y situées sur ce cours d'eau sont considérées comme voisines. Ceci est illustré sur la figure 3.

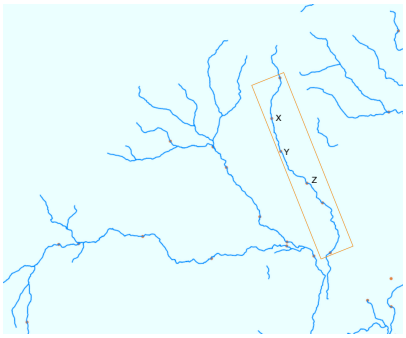

FIG. 3 – *Zonage par la méthode du cours d'eau*

2. Le *k-voisinage* : découpage de l'espace en zones autour de chaque station en exploitant les coordonnées x et y. Dans chacune de ces zones, nous avons regroupé des voisins qui se trouvent à l'intérieur d'une zone de k km^2 centrée sur une station X. Dans ce type de découpage, nous ne prenons pas en compte le cours d'eau comme illustrée sur la figure 4.

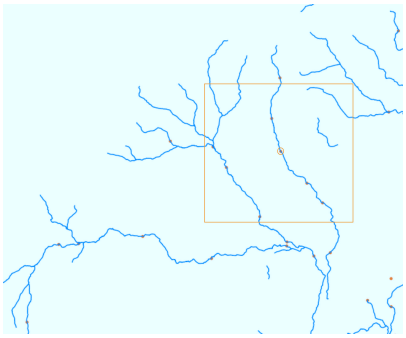

FIG. 4 – *Zonage par la méthode du k-voisinage*

Grace à ces deux méthodes de zonage, nous sommes capables de rapprocher les stations dans des zones et ainsi, de regrouper les données pour le critère de comptage de la fréquence

des séquences. Le calcul ne se fera plus par station mais par zone ainsi définie, ceci permet d'obtenir des motifs séquentiels plus pertinents vis-à-vis du caractère hétérogène des relevés. Nous détaillons la méthode d'extraction dans la section suivante. Ces deux découpages nous permettent de mettre en évidence deux hypothèses différentes pour l'influence de la pollution. (i) La première hypothèse est que la pollution mesurée dans un cours d'eau donné, à une station particulière X va d'une part avoir potentiellement un impact sur les stations situées en aval de X et d'autre part, que l'origine de la pollution est liée à un phénomène situé en amont de X. Le découpage de type "cours d'eau" nous permet donc d'évaluer cette hypothèse en "moyennant" les indicateurs de pollution à l'intégralité d'un cours d'eau. (ii) La seconde hypothèse est que la pollution mesurée dans une station X est le résultat d'une pollution dont l'origine peut se situer sur le même cours d'eau, dans des nappes souterraines, dans des zones agricoles voisines, etc. Le découpage en "k-voisinage" nous permet donc de moyenner les indicateurs de pollution dans des zones suffisamment grandes pour observer des effets potentiellement non locaux au cours d'eau associé à la station X.

4.2 La méthode d'extraction de séquences

Nous avons adopté l'algorithme *prefixspan* (Pei et al., 2004) pour l'extraction de motifs séquentiels pour son efficacité dans la fouille de gros volumes de données. Cette méthode adopte une stratégie "diviser pour régner" en réalisant un parcours en profondeur des données accompagné de projections successives de la base de données. Une projection de la base de données selon une séquence s est constituée de l'ensemble des suffixes des séquences présentes dans la base de données préfixées par s. L'objectif est de réduire le parcours de l'espace de recherche. Pour cela, *prefixspan* propose d'analyser les préfixes communs que présentent les séquences de données de la base à traiter. À partir de cette analyse, l'algorithme construit des bases de données intermédiaires qui sont des projections de la base d'origine déduites à partir des préfixes identifiés. Ensuite, dans chaque base obtenue, *prefixspan* cherche à faire croître la taille des motifs séquentiels découverts en appliquant la même méthode de manière récursive.
Exemple : Soient la base de séquences décrite dans le tableau 3 et un support minimal $minSupp = 0,5$ pour lequel nous appliquons la méthode *prefixpan* :

Id	Séquence
1	$< (a,b)(c) >$
2	$< (a,b)(a,b) >$
3	$< (a,e)(d,e) >$

TAB. 3 – *Base de séquences*

Les items fréquents dans cette base sont : $< a >$ avec un support de 3 et $< b >$ avec un support de 2. $< c >$, $< d >$ et $< e >$ ne sont pas fréquents. Les projections de ces deux items sur la base initiale de séquences sont présentées dans le tableau 4 :

Finalement, les motifs fréquents (apparaissant au moins 2 fois dans la base de séquences initiale) sont : $< a >$, $< b >$ et $< (a,b) >$.

Pour réaliser les expérimentations, nous avons utilisé le Framework *SPMF (Sequential Pattern Mining Framework)* implementé par Phillipe Fournier-Viguer et disponible sur `http://www.philippe-fournier-viger.com/spmf/`. Nous obtenons des en-

Préfixe	Base projetée	Motifs fréquents
$< a >$	$< (\_, b)(c) >$ $< (\_, b)(a, b) >$ $< (\_, e)(d, e) >$	$< (a, b) >: 2$
$< b >$	$< (c) >$ $< (a, b) >$	

TAB. 4 – *Projections du tableau de séquences pour les items* $< a >$ *et* $< b >$

sembles de motifs très nombreux que nous filtrons grâce à la mesure de moindre contradiction temporelle présentée dans la section suivante 4.3.

4.3 La mesure de moindre contradiction temporelle

Dans le domaine de l'extraction de connaissances, il est courant d'obtenir des séquences fréquentes plus nombreuses que les données initiales. Choisir les séquences les plus intéressantes reste un problème difficile, souvent étroitement lié aux données manipulées.

4.3.1 Mesures de qualité : bref état de l'art et motivations

Dans le cadre plus général des règles d'association de la forme antécédent $\rightarrow$ conséquent, de nombreuses mesures de qualité ont été proposées afin d'évaluer l'intérêt des règles extraites et donc de réduire l'ensemble des règles qu'un expert humain devra analyser. Une étude comparative d'un large ensemble de ces mesures de qualité est présenté dans (Tan et al., 2002). Parmi les critères d'intérêt recherchés dans les règles d'association, les plus fréquemment utilisés évaluent l'écart à l'indépendance entre l'antécédent et le conséquent de la règle. Cette évaluation peut être réalisée à l'aide de nombreuses mesures de qualité telles que : la confiance (Agrawal et Srikant, 1995), l'évaluation du faible nombre de contre-exemples associés à la règle (Azé, 2003), la notion d'étonnement statistique (Lerman et Guillaume, 2011; Gras et al., 2001; Lerman et Azé, 2007; Lerman et al., 1981), *etc.* Les mesures existantes ne peuvent s'appliquer que dans le cadre de règles de la forme antécédent $\rightarrow$ conséquent. Or, dans notre contexte d'étude, nous n'avons pas cette notion de règle mais seulement la notion de séquence d'itemsets fréquents. À l'exception du support, les mesures de qualité existantes ne peuvent donc pas être utilisées directement.

Bien que le domaine de recherche centré sur les mesures de qualité soit très actif (Jalali-Heravi et Zaïane, 2010), à notre connaissance, il n'existe aucune approche permettant de prendre en compte les aspects temporels. Les mesures de qualité se basent généralement sur des formules faisant intervenir le support (Geng et Hamilton, 2006) mais ne se soucient pas de l'ordre inhérent dans la séquence associé à l'axe temporel.

Nous avons ainsi choisi de nous focaliser sur la recherche de séquences fréquentes qui sont peu contredites par les données. Pour cela, nous proposons d'étendre la mesure de moindre contradiction (MC), définie par (Azé, 2003) dans le cadre des règles d'association, au contexte des séquences temporelles. Rappelons que cette mesure, dans le cadre d'une règle d'association $A \rightarrow B$ où A et B sont deux ensembles d'items disjoints, est définie par :

$$MC(A \rightarrow B) = \frac{supp(AB) - supp(A\overline{B})}{supp(B)} \quad (1)$$

où $A\overline{B}$ est l'itemset tel que A est présent et B absent.

Nous proposons donc l'extension de cette mesure au cas de séquences temporelles afin de pouvoir évaluer l'influence de la temporalité sur les motifs extraits.

Nous avons choisi d'étendre la moindre contradiction aux séquences temporelles pour deux raisons principales. Premièrement, cette mesure est simple à comprendre et à mettre en œuvre. Elle est donc relativement simple à appréhender par les experts. Deuxièmement, des précédents travaux ont exhibés des comportements intéressants de cette mesure pour extraire des pépites de connaissances (Azé, 2003), pour résister au bruit (Azé et al., 2007), *etc*. Enfin, d'autres mesures pourront également être étendues aux séquences temporelles comme par exemple le lift dont la définition est "proche" de celle de MC.

4.3.2 Définition de la mesure de moindre contraction temporelle

Soit la séquence fréquente S, la *moindre contradiction temporelle* de S, notée $MCT(S)$ est définie par :

$$MCT(S) = \frac{supp(S) - \sum\limits_{s_d \in S_d} supp(s_d)}{\sum\limits_{s_t \in S_t} supp(s_t)} \tag{2}$$

où $\begin{cases} S_d & \text{l'ensemble des séquences incluant tous les itemsets de la séquence } S \\ & \text{mais dans un ordre différent} \\ S_t & \text{l'ensemble des séquences incluant tous les items} \\ & \text{qui sont apparus dans la séquence } S \end{cases}$

Cette extension de la moindre contradiction permet de conserver l'esprit initial de la mesure qui vise à évaluer le nombre de fois où une règle est vérifiée *vs* le nombre de fois où elle est invalidée. Une règle qui est plus fréquemment vérifiée qu'invalidée est a priori intéressante. Comme pour la version "classique", cette mesure est normalisée. Ici, la normalisation est effectuée par rapport à la somme des supports des séquences pouvant être construites à partir des items composant la séquence d'intérêt.

Exemple :
Soient les séquences et leur support :

$$\begin{cases} S_1 =< (AB)(BC) >, supp(S_1) = 0,25 \\ S_2 =< (BC)(AB) >, supp(S_2) = 0,10 \\ S_3 =< (AB)(CE) >, supp(S_3) = 0,12 \\ S_4 =< (AB) >, supp(S_4) = 0,13 \\ S_5 =< (EA)(BC) >, supp(S_5) = 0,20 \end{cases}$$

Alors,

$$MCT(S_1 =< (AB)(BC) >) = \frac{supp(S_1) - \sum\limits_{s_d \in S_d} supp(s_d)}{\sum\limits_{s_t \in S_t} supp(s_t)} = \frac{0,25 - 0,10}{0,67} = 0,224$$

$$avec \begin{cases} supp(S_1) = 0,25 \\ S_d = \{S_2\} \\ S_t = \{S_1, S_2, S_3, S_5\} \end{cases}$$

On retrouve (BC) et (AB) dans S_2 (qui a les mêmes itemsets que la séquence S_1 mais dans un ordre différent) et on retrouve les items A, B et C dans S_1, S_2, S_3 et S_5, mais pas dans S_4 qui ne contient que les items A et B.

4.3.3 Algorithme

L'algorithme 1 décrit les étapes que nous proposons pour le calcul de *la moindre contradiction temporelle*. Cet algorithme se déroule en deux étapes : (1) Tout d'abord, la recherche de séquences contenant des itemsets communs entre la séquence traitée et la séquence candidate sans considérer l'ordre d'apparition (première boucle interne) ; (2) la recherche de tous les items appartenant à la séquence traitée dans la séquence candidate (deuxième boucle interne). Cet algorithme a une complexité d'au plus $O(n^2)$ où n représente le nombre de n-uplets de la base de données.

4.4 Le regroupement de séquences

Dans le domaine de l'extraction de connaissances, il est souvent nécessaire de comparer les objets extraits de façon à trouver des régularités ou de construire des classes d'objets homogènes. Dans cet article, nous utilisons la mesure de similarité S^2MP proposée dans (Saneifar et al., 2008) qui prend en compte les caractéristiques et la sémantique des motifs séquentiels. En effet, cette mesure permet de mesurer la proximité des motifs trouvés afin de proposer des regroupements cohérents. S^2MP est une mesure de similarité permettant de comparer deux séquences. Elle est basée sur deux scores : le score d'association qui prend en compte le nombre d'éléments communs entre deux séquences et le score d'ordre qui prend en compte l'ordre commun de ces éléments.

Le score d'association S_m : On associe tous les itemsets de la première séquence avec tous les itemsets de la seconde séquence et pour chaque association, on calcule un poids (nombre d'items communs divisé par le nombre d'items des deux itemsets comparés). Pour chaque combinaison possible d'associations, on calcule une moyenne des poids et on conserve la combinaison associée à la meilleure moyenne.

Exemple :

Soit deux séquences $S_1 =< (ABC)(AB)(CD) >$ et $S_2 =< (AB)(CA)(A) >$. Le poids associé à l'association entre les itemsets (ABC) dans S_1 et (AB) dans S_2 est égal à $2/((3+2)/2) = 0.8$. De la même manière, on associe (ABC) et (CA) avec un score de 0.8 et (ABC) et (A) avec un score de 0.5. Pour le premier itemset de S_1 (ABC), l'association retenue sera la première, soit (ABC) avec (AB) avec le score de 0.8.

On procède de même avec les autres itemsets de S_1 pour associer (AB) et (A) avec un poids de 0.6 ainsi que (CD) et (CA) avec un poids de 0.5.

Finalement, S_m correspond à la moyenne des poids de ces 3 associations, soit 0.65.

Entrées : mBD : Base de données de motifs fréquents et leurs supports
Sorties : *la moindre contradiction temporelle* de chaque séquence $S \in mBD$
début
  vecteur MCT ;
  pour tous les *(séquence* $S_1 \in mBD$*)* **faire**
    float $suppS_d \leftarrow 0$;
    float $suppS_t \leftarrow 0$;
    pour tous les *(séquence* $S_2 \in mBD - \{S_1\}$*)* **faire**
      boolean $all\_in \leftarrow vrai$;
      tant que *((all_in)* **et** *(*$\forall$ *itemsets* $IS \in S_1$*))* **faire**
        si *(*$IS \not\subseteq S_2$*)* **alors**
          $all\_in \leftarrow faux$;
        sinon
          prochain $IS \in S_1$;
        fin
      fin
      si *(*all_in*)* **alors**
        $suppS_d \leftarrow suppS_d + supp(S_2)$;
      fin
      $all\_in \leftarrow vrai$;
      tant que *((all_in)* **et** *(*$\forall$ *item* $I \in S_1$*))* **faire**
        si *(*$I \not\subset S_2$*)* **alors**
          $all\_in \leftarrow faux$;
        sinon
          prochain $I \in S_1$;
        fin
      fin
      si *(*all_in*)* **alors**
        $suppS_t \leftarrow suppS_t + supp(S_2)$;
      fin
    fin
    $MCT(S_1) \leftarrow \frac{supp(S_1) - suppS_d}{suppS_t}$;
  fin
  retourner MCT;
fin

Algorithme 1 : Calcul de la moindre contradiction temporelle

Le score d'ordre S_o : Pour calculer ce score, on agrège 2 sous-scores $TotalOrder$, le pourcentage d'associations respectant l'ordre de la séquence et $PositionOrder$, correspondant à l'écart entre deux associations consécutives. Pour cela, on utilise la formule :

$$S_o = max\left\{totalOrder(sub) * positionOrder(sub)\right\} \quad (3)$$

$$avec \text{ sub} \in \{ \text{ sous-séquences croissantes de la séquence initiale} \}.$$

Exemple :
l'ordre des itemsets associés à S_1 dans S_2 est (1,3,2). On trouve 2 sous-séquences croissantes (1,3) et (1,2).

$$\begin{aligned} TotalOrder &= 2/3 \\ PositionOrder(\{1,3\}) &= 1-(1-2)/3 = 2/3 \\ PositionOrder(\{1,3\}) &= 1-(2-1)/3 = 2/3 \\ S_o &= max(\frac{2}{3} * \frac{2}{3}; \frac{2}{3} * \frac{2}{3}) \\ &= 0,44 \end{aligned}$$

Le calcul de similarité S^2MP est le produit de ces deux scores : le score de mapping et le score d'ordre.

$$\begin{aligned} S^2MP &= S_m * S_o/2 \\ &= 0,63 * 0,44/2 \\ &= 0,53 \end{aligned}$$

Cette mesure a été utilisée pour comparer les motifs (voir section 4.1) que nous présentons aux experts.

5 Expérimentations

Les expérimentations ont été réalisées à partir d'une base de données d'indicateurs biologiques relevés dans les rivières de la Saône (voir figure 1). Ce jeu de données est constitué de 10 attributs d'analyse associés à des caractéristiques biologiques de l'eau (e.g. l'indice biologique global normalisé IBGN et l'indice biologique diatomée IBD). Le code de la station de prélèvement (*codstace*) et la date de prélèvement des échantillons (*rdate*) sont aussi des attributs stockés dans la base de données. Au total, le jeu de données est constitué de 12 caractéristiques et $2\,534$ lignes. Le tableau 5 décrit l'ensemble des attributs A_i ainsi que leur domaine de valeur $dom(A_i)$.

5.1 La discrétisation des données

L'étude de la répartition des valeurs pour l'ensemble des attributs, nous a permis de définir une méthode de discrétisation. Dans notre cas, les composantes sont bien séparées, le nombre d'observations suffisant et la seule considération de l'histogramme des fréquences (voir figures

A_i	**dom**(A_i)
codstace	[6000850 . . . 6940940]
rdate	01/04/1993 . . . 16/10/2008
ibgn	[0, 1, . . . 20, -100]
gr_indic	[0, 1, . . . 9, -100]
var_taxo	[2 . . . 59, -100]
ibgn_etat	{BE, Emauv, Emedio, Emoy, ND, ND_No_Ref, ND_No_Type, No_Ref, No_Type, TBE}
ibgn_note	[0, 1, . . . 4, -100, -101, . . . -104]
ibd	[4.6, 6.0, . . . 20.0, -100]
ibd2007	[5.9, 6.1, . . . 20.0, -100]
ibd_etat	{BE, Emauv, Emedio, Emoy, ND, ND_No_Ref, ND_No_Type, No_Ref, No_Type, TBE}
ibd_notev	[0, 1, . . . 4, -100, -102, -104]
ibgn_ibd	[0, 1, 2, 3, -1, -2, -3, -100]

TAB. 5 – *Descriptions des attributs A_i et leur domaine*

5 et 6) fourni une estimation correcte du nombre de constituants et de leurs valeurs. Les descripteurs *gr_indic*, *ibgn* et *var_taxo* suivent une distribution quasi normale (voir par exemple la figure 5) alors que les descripteurs *ibd* et *ibd_2007* suivent des distributions très différentes (voir par exemple la figure 6).

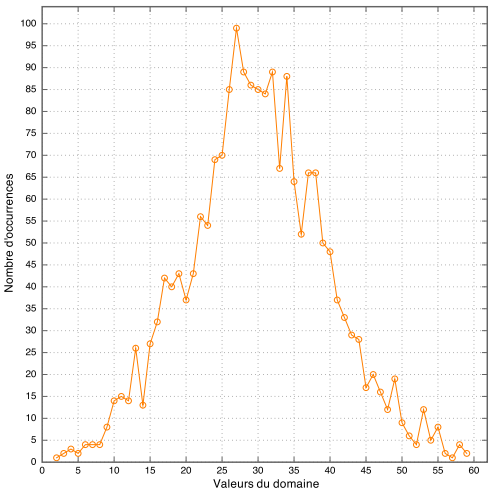

FIG. 5 – *Histogramme pour l'attribut var_taxo*

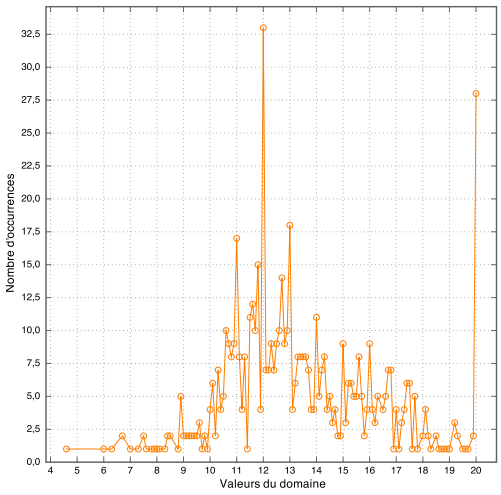

FIG. 6 – *Histogramme pour l'attribut ibd2007*

Pour réaliser la discrétisation, nous avons choisi une répartition à *fréquences égales*. Pour cela, nous avons construit la courbe de fréquences cumulées et analysé cette courbe pour déterminer les limites de chaque classe en fonction de seuils observés sur la courbe. Les plages de valeurs sont choisies pour égaliser la distribution des valeurs de la variable donnée (par exemple *var_taxo* dans la figure 5). Avec ce type de discrétisation, nous obtenons des plages de valeurs équilibrées.

Les valeurs discrétisées associées au domaine des attributs d'analyse sont présentées dans le tableau 6.

A_i	**Discrétisation**
ibgn	[0 . . . 10] , [11 . . . 15], [16 . . . 20]
gr_indic	[0 . . . 4], [5 . . . 6] , [7 . . . 9]
var_taxo	[2 . . . 20], [21 . . . 30] , [31 . . . 40], [41 . . . 59]
ibgn_etat	{TBE, BE, Emauv, Emedio, Emoy}
ibgn_note	{0, 1, 2, 3, 4}
ibd	[4 . . . 10],]11 . . . 13],]14 . . . 16],]17 . . . 20]
ibd2007	[5 . . . 10],]11 . . . 13],]14 . . . 16],]17 . . . 20]
ibd_etat	{TBE, BE, Emauv, Emedio, Emoy}
ibd_notev	{0, 1, 2, 3, 4}
ibgn_ibd	{0, 1, 2, 3}

TAB. 6 – *Discrétisation des valeurs du domaine* $dom(A_i)$

5.2 La spatialisation des données

Nous avons extrait des séquences de motifs selon trois approches de spatialisation (voir section 4.1) :

1. En considérant l'espace comme une seule unité, c'est-à-dire, sans le découper en zones, *NZ* ;
2. En utilisant le voisinage du type *cours d'eau*, pour diviser l'espace en zones plus ou moins hétérogènes : le découpage de l'espace en utilisant ce type de voisinage nous a fourni un total de 233 zones ;
3. En considérant le *k-voisinage*, où *k* a été fixé dans un premier temps à 100, définissant ainsi des zones de 10 kms $\times$ 10 kms centrées sur une station d'intérêt à $10\,000$ kilomètres : nous avons obtenu 223 zones selon cette approche.

5.3 L'extraction de motifs séquentiels

Les premiers résultats des expérimentations appliquées aux données avec les trois approches de spatialisation proposées sont :

1. Pour la première approche, *NZ*, nous avons appliqué la méthode choisie à un jeu de données composé de 711 stations de prelèvement identifiées par le code de la station et utilisé un support minimal de $0,3$. Nous avons obtenu 22 motifs fréquents, tous de taille 1. Le tableau 7 montre quelques motifs extraits ;

Motifs	**Fréquence**	**Support**
<(ibgn_etat_TBE)>	228/711	0,32
<(ibgn_etat_TBE, ibgn_note_4)>	228/711	0,32
<(ibgn_0-10, gr_indic_0-4)>	224/711	0,32
<(ibgn_etat_BE, ibgn_note_3)>	217/711	0,31
. . .	. . .	. . .

TAB. 7 – *Exemple de résultats obtenus pour les données de la première approche*

2. Pour la deuxième approche, *cours d'eau*, nous avons appliqué l'algorithme choisi à un jeu de données composé de 233 zones avec un support minimal de $0,3$. Nous avons obtenu 564 motifs fréquents, parmi lesquels 110 sont de taille 1, 361 sont de taille 2, 90 de taille 3 et 3 de taille 4. Une partie des résultats trouvés est présentée dans le tableau 8 ;

Motifs	Fréquence	Support
<(gr_indic_5-6, var_taxo_21-30, ibgn_etat_Emoy, ibgn_note_2)>	71/233	0,3
<(ibgn_note_2) (ibgn_etat_Emoy, ibgn_note_2)>	72/233	0,31
<(ibgn_11-15, ibgn_etat_Emoy, ibgn_note_2) (ibgn_11-15)>	81/233	0,35
<(ibgn_11-15) (ibgn_11-15, var_taxo_21-30)>	95/233	0,41
<(var_taxo_21-30) (var_taxo_21-30) (ibgn_11-15)>	84/233	0,36
<(var_taxo_21-30) (ibgn_11-15, var_taxo_21-30) (var_taxo_21-30)>	77/233	0,33
<(ibgn_11-15, var_taxo_21-30) (ibgn_11-15) (ibgn_11-15, var_taxo_21-30)>	72/233	0,31
<(ibgn_11-15) (ibgn_11-15, var_taxo_21-30) (ibgn_11-15, var_taxo_21-30)>	70/233	0,3
<(var_taxo_21-30) (var_taxo_21-30) (ibgn_11-15) (ibgn_11-15)>	70/233	0,3
<(var_taxo_21-30) (ibgn_11-15) (ibgn_11-15) (ibgn_11-15)>	70/233	0,31
...	...	...

TAB. 8 – *Exemple de résultats obtenus pour les données de la deuxième approche*

3. Pour la troisième approche, *k-voisinage*, nous avons appliqué la méthode choisie à un jeu de données composé de 223 zones avec un support minimal de $0,3$. Nous avons obtenu 138 motifs fréquents de taille 1, 1 174 motifs fréquents de taille 2, 658 de taille 3, 104 de taille 4 et 8 motifs de taille 5. Au total, 2 082 motifs fréquents sont extraits. Une partie des résultats trouvés est présentée dans le tableau 9.

Motifs	Fréquence	Support
<(var_taxo_21-30, ibgn_etat_Emoy)>	108/223	0,48
<(ibgn_16-20, gr_indic_7-9, var_taxo_31-40, ibgn_etat_TBE, ibgn_note_4)>	85/223	0,38
<(ibgn_note_2) (ibgn_etat_Emoy, ibgn_note_2)>	81/223	0,36
<(var_taxo_21-30, ibgn_note_2) (ibgn_11-15, var_taxo_21-30)>	79/223	0,35
<(gr_indic_7-9) (ibgn_16-20, gr_indic_7-9, ibgn_etat_TBE, ibgn_note_4)>	87/223	0,39
<(ibgn_etat_Emoy, ibgn_note_2) (var_taxo_21-30)>	93/223	0,42
<(var_taxo_21-30, ibgn_etat_Emoy) (var_taxo_21-30) (ibgn_11-15)>	70/223	0,31
<(var_taxo_21-30, ibgn_etat_Emoy, ibgn_note_2) (ibgn_11-15) (ibgn_11-15, var_taxo_21-30)>	70/223	0,31
<(var_taxo_21-30) (var_taxo_21-30) (ibgn_11-15)>	94/223	0.42
<(ibgn_11-15, var_taxo_21-30) (gr_indic_5-6) (var_taxo_31-40)>	67/223	0,3
<(ibgn_11-15, var_taxo_21-30) (var_taxo_31-40) (ibgn_11-15, var_taxo_21-30)>	67/223	0,3
<(ibgn_11-15, var_taxo_21-30) (ibgn_11-15) (ibgn_11-15) (var_taxo_31-40)>	69/223	0,31
<(ibgn_11-15, var_taxo_21-30) (ibgn_11-15) (ibgn_11-15) (ibgn_11-15) (ibgn_11-15)>	67/223	0,3
...	...	...

TAB. 9 – *Exemple de résultats obtenus pour les données de la troisième approche*

Le nombre de séquences obtenues lors de l'exécution de l'algorithme *prefixspan* sur le jeu de données avec les trois approches de spatialisation et un support minimal de $0,3$ est respectivement de 22 sans zonage, de 564 pour le zonage *cours d'eau* et de 2 082 pour le zonage *k-voisinage*.

Il est intéressant de constater que nous obtenons peu de motifs pour la première approche à la différence des deux autres approches de spatialisation.

5.4 Une mesure objective de validation : la moindre contradiction temporelle

Nous avons appliqué une mesure objective (non dépendante d'un point de vue mais associée à un calcul) de validation sur les connaissances extraites pour filtrer les motifs. En effet, même si en terme de volume sur le jeu de données traité, une validation exhaustive est envisageable, cela ne le sera plus dès lors que le jeu de données sera étendu au niveau national.

La moindre contradiction temporelle :
La *mesure de moindre contradiction temporelle MCT* (voir section 4.3.2) a été calculée de la façon suivante :

Soit mBD la base de données des motifs séquentiels obtenus après l'exécution de l'algorithme *prefixspan* (Pei et al., 2004) sur le jeu de données du bassin versant de la Saône en considérant la spatialisation du *plus proche voisin*, par exemple. Soit une séquence $S \in mBD$ et son support $supp(S)$ décrits dans le tableau suivant :

Séquence	Support
$< (ibgn\_16\text{-}20, ibgn\_etat\_TBE)(var\_taxo\_31\text{-}40) >$	$0,34$

Pour le calcul de S_d, nous cherchons les itemsets $(ibgn\_16\text{-}20, ibgn\_etat\_TBE)$ et $(var\_taxo\_31\text{-}40)$ dans toutes les séquences S de la base de données sans considérer la position d'apparition dans S. Nous les trouvons deux fois :

Séquence	Support
$< (ibgn\_16\text{-}20, ibgn\_etat\_TBE)(ibgn\_11\text{-}15)(var\_taxo\_31\text{-}40) >$	$0,34$
$< (var\_taxo\_31\text{-}40)(ibgn\_16\text{-}20, ibgn\_etat\_TBE) >$	$0,32$

Finalement, la valeur S_d pour la séquence *<(ibgn_16-20, ibgn_etat_TBE) (var_taxo_31-40)>* est $0,66$.

Le calcul de S_t se fait de façon analogue au calcul de S_d. Nous cherchons les *items* appartenant à la séquence $< (ibgn\_16\text{-}20, ibgn\_etat\_TBE)(var\_taxo\_31\text{-}40) >$ dans toutes les séquences S de la base de données des motifs fréquents mBD. Nous avons trouvé ces *items* dans les séquences suivantes :

Séquence	Support
$< (ibgn\_16\text{-}20, gr\_indic\_7\text{-}9, var\_taxo\_31\text{-}40, ibgn\_etat\_TBE) >$	$0,34$
$< (ibgn\_16\text{-}20, gr\_indic\_7\text{-}9, var\_taxo\_31\text{-}40, ibgn\_etat\_TBE, ibgn\_note\_4) >$	$0,34$
$< (ibgn\_16\text{-}20, var\_taxo\_31\text{-}40, ibgn\_etat\_TBE) >$	$0,36$
$< (ibgn\_16\text{-}20, var\_taxo\_31\text{-}40, ibgn\_etat\_TBE, ibgn\_note\_4) >$	$0,36$
$< (ibgn\_16\text{-}20, gr\_indic\_7\text{-}9, ibgn\_etat\_TBE, ibgn\_note\_4)(var\_taxo\_31\text{-}40) >$	$0,31$
...	...

La somme des supports des séquences $s_t \in S_t$ est égale à $3,34$.

Finalement, *la moindre contradiction temporelle (MCT)* pour la séquence *<(ibgn_16-20, ibgn_etat_TBE) (var_taxo_31-40)>* est :

$$MCT(< (ibgn\_16\text{-}20, ibgn\_etat\_TBE)(var\_taxo\_31\text{-}40) >) = \frac{0,34 - 0,66}{3,34}$$
$$= -0,09580838323353$$

La mesure objective d'évaluation *MCT* a été appliquée aux motifs obtenus lors de l'exécution de l'algorithme choisi sur le jeu de données des bassins versants de la Saône pour les trois approches spatiales proposées.

Les tableaux 10, 11 et 12 montrent quelques séquences et les valeurs associées à S_d, S_t et la valeur de la *moindre contradiction temporelle MCT* pour les différentes approches de spatialisation :

Séquence	**Support**	S_b	S_t	**MCT**
<(ibgn_etat_TBE, ibgn_note_4)	0,32	0,32	4,32	2,0
<(ibgn_11-15, var_taxo_21-30)	0,39	0,39	0,39	2,0
<(var_taxo_21-30)	0,5	0,5	0,89	1,1236
<(ibgn_0-10)	0,36	0,36	0,68	1,05882
<(ibgn_note_2)	0,32	0,32	0,64	1,0
...	...	...	...	...

TAB. 10 – *MCT pour les données sans zonage NZ*

Séquence	**Support**	S_b	S_t	**MCT**
<(var_taxo_21-30) (ibgn_16-20, var_taxo_31-40)>	0,3	0,3	0,3	2,0
<(ibgn_0-10, gr_indic_0-4, ibgn_etat_Emedio, ibgn_note_1)>	0,32	0,32	0,32	2,0
<(ibgn_0-10, ibgn_etat_Emedio, ibgn_note_1)>	0,34	0,34	0,66	1,0303
<(ibgn_note_1)	0,35	0,35	2,68	0,261194
<(ibgn_note_4) (ibgn_etat_TBE)>	0,34	0,68	27,7	0,0368231
<(var_taxo_21-30) (var_taxo_21-30)(ibgn_11-15)>	0,3	0,3	32,08	0,0187032
<(ibgn_etat_TBE) (ibgn_etat_TBE)>	0,34	0,34	42,02	0,0161828
<(ibgn_11-15)(ibgn_11-15) (ibgn_11-15)>	0,32	0,32	69,68	0,00918484
...	...	...	...	...

TAB. 11 – *MCT pour les données en considérant le zonage cours d'eau*

Séquence	**Support**	S_b	S_t	**MCT**
<(ibgn_11-15) (ibgn_16-20, gr_indic_7-9)>	0,33	0,33	0,33	2,0
<(var_taxo_21-30) (ibgn_etat_TBE, ibgn_note_4)>	0,33	0,33	0,64	1,03125
<(gr_indic_7-9) (ibgn_11-15, var_taxo_21-30)>	0,36	0,72	1,06	1,01887
<(gr_indic_7-9) (ibgn_etat_BE, ibgn_note_3)>	0,31	0,63	0,97	0,969072
<(var_taxo_21-30) (var_taxo_31-40)>	0,42	1,73	2,74	0,784671
<(ibgn_etat_TBE, ibgn_note_4) (var_taxo_31-40, ibgn_note_4)>	0,31	0,31	6,59	0,0940819
<(gr_indic_7-9) (gr_indic_7-9)>	0,3	1,67	49,97	0,0394236
<(var_taxo_21-30) (var_taxo_21-30) (ibgn_11-15)>	0,3	0,3	32,08	0,0187032
<(ibgn_11-15) (ibgn_11-15) (ibgn_11-15)>	0,32	0,32	69,68	0,00918484
...	...	...	...	...

TAB. 12 – *MCT pour les données en considérant le zonage k-voisinage*

5.5 Comparaison de séquences et restitution

Si la méthode de la moindre contradiction temporelle permet de filtrer les motifs, elle ne permet pas de guider l'exploration de l'expert dans l'analyse des motifs retenus. A partir d'un ensemble de séquences d'intérêt défini par les experts, nous utilisons la mesure S^2MP (voir sous-section 4.4) pour comparer les séquences au niveau des itemsets et de leur position dans la séquence ainsi qu'au niveau de la ressemblance des items dans les itemsets.

Pour le jeu de données sans zonage, nous avons utilisé comme ensemble initial les 3 séquences présentées dans le tableau 13, que nous avons comparé à toutes les autres séquences grâce à la mesure S^2MP.

Pour le jeu de données associé au zonage *cours d'eau*, composé de 564 tuples, les séquences choisies pour être comparées aux autres sont indiquées dans le tableau 14 :

Séquences
(ibgn_note_2)
(ibgn_note_3)
(ibgn_note_4)

TAB. 13 – *Séquences d'intérêt pour des données sans zonage*

Séquences
(var_taxo_21-30)
(ibgn_16-20, var_taxo_31-40, ibgn_etat_TBE)
(ibgn_11-15, ibgn_etat_Emoy)
(ibgn_11-15, gr_indic_7-9)
(gr_indic_7-9)(ibgn_16-20, var_taxo_31-40)
(ibgn_etat_BE, ibgn_note_3)(ibgn_note_3)
(ibgn_11-15, ibgn_note_2)(ibgn_11-15, var_taxo_21-30)
(ibgn_11-15, gr_indic_5-6)(var_taxo_21-30)
(ibgn_11-15, var_taxo_21-30)(ibgn_note_2)
(ibgn_11-15, var_taxo_21-30)(ibgn_etat_BE)

TAB. 14 – *Séquences d'intérêt associées au zonage cours d'eau*

Pour chacune de ces séquences d'intérêt, nous avons retenu les plus similaires grâce à la mesure S^2MP. Par exemple, pour la séquence *(var_taxo_21-30)*, nous obtenons, les séquences du tableau 15 :

Séquences	Distance
(var_taxo_21-30, ibgn_etat_Emoy)	0,9
(gr_indic_5-6)	0,75
(gr_indic_5-6, var_taxo_21-30)	0,9
(gr_indic_5-6, var_taxo_21-30, ibgn_etat_Emoy)	0,833333
(ibgn_11-15, var_taxo_21-30)	0,9
...	...

TAB. 15 – *Séquences et leur distance à la séquence (var_taxo_21-30)*

De nombreuses séquences ont la même distance aux séquences d'intérêt. Cette caractéristique est propre à notre jeu de données en raison de la grande quantité de données manquantes et à la faible taille des séquences extraites ainsi qu'à la mesure choisie pour les comparer. Le tableau 16 indique par séquence d'intérêt, le nombre de séquences associées à une même valeur de ressemblance.

Ainsi, pour la séquence d'intérêt *(var_taxo_21-30)*, il y a 16 séquences ayant une distance de ressemblance de $0,833333$. La figure 7 représente les séquences associées à une séquence d'intérêt et leur distance.

Pour notre jeu de données associé au zonage *k-voisin* et composé de $2\,082$ tuples, nous avons retenu 10 séquences d'intérêt. Nous avons réalisé le même calcul que pour le jeu de données associé au zonage *cours d'eau*, pour obtenir les résultats du tableau 17. Ce tableau montre que la séquence *(gr_indic_5-6, ibgn_etat_Emoy, ibgn_note_2)* a 52 séquences avec une distance de ressemblance de 0,75. La figure 8 montre les séquences regroupées par séquence d'intérêt et leur distance.

Séquences d'intérêt	Distance	Coïncidences
(var_taxo_21-30)	0,0	4
(var_taxo_21-30)	0,75	1
(var_taxo_21-30)	0,833333	16
(var_taxo_21-30)	0,9	5
(var_taxo_21-30)	1,0	2

TAB. 16 – *Nombre de coïncidences par distance pour la séquence (var_taxo_21-30)*

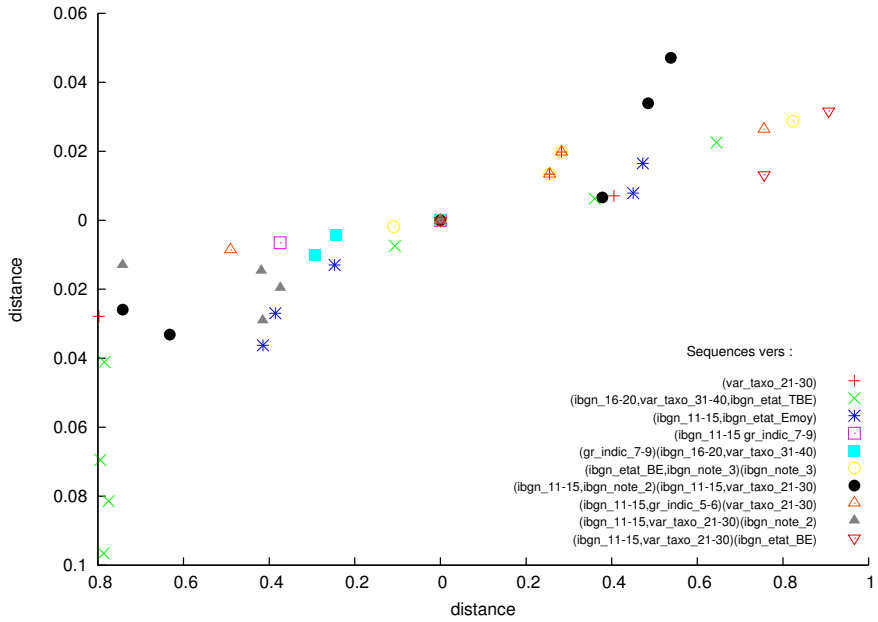

FIG. 7 – *Distance des séquences à la séquence d'intérêt pour le zonage cours d'eau*

La figure 9 montre les séquences regroupées par séquence et leur distance pour les deux approches i.e. zonage *cours d'eau* et le zonage *k-voisin*.

6 Conclusion

Dans cet article, nous avons présenté les premières étapes d'un projet de fouille de données hydrologiques. Nous avons plus particulièrement appliqué un algorithme classique d'extraction de motifs séquentiels selon trois approches de spatialisation. Nous avons souligné les problèmes qui sont posés selon les choix réalisés en matière de spatialisation et leur influence sur le nombre de motifs extraits. Nous avons proposé une mesure de validation objective : la *moindre contradiction temporelle* qui permet de donner aux experts une mesure adaptée pour l'évaluation des motifs obtenus. Nous avons également appliqué une mesure de similarité pour comparer des séquences de motifs extraites selon les différentes proximités spatiales. Ces travaux ont été menés en "aveugle" c'est-à-dire sans intervention des spécialistes des données.

Séquences d'intérêt	Distance	Coïncidences
(gr_indic_5-6,ibgn_etat_Emoy,ibgn_note_2)	0,0	5
(gr_indic_5-6,ibgn_etat_Emoy,ibgn_note_2)	0,75	52
(gr_indic_5-6,ibgn_etat_Emoy,ibgn_note_2)	0,833333	168
(gr_indic_5-6,ibgn_etat_Emoy,ibgn_note_2)	0,9	1

TAB. 17 – *Nombre de coïncidences par distance pour la séquence (gr_indic_5-6, ibgn_etat_Emoy, ibgn_note_2)*

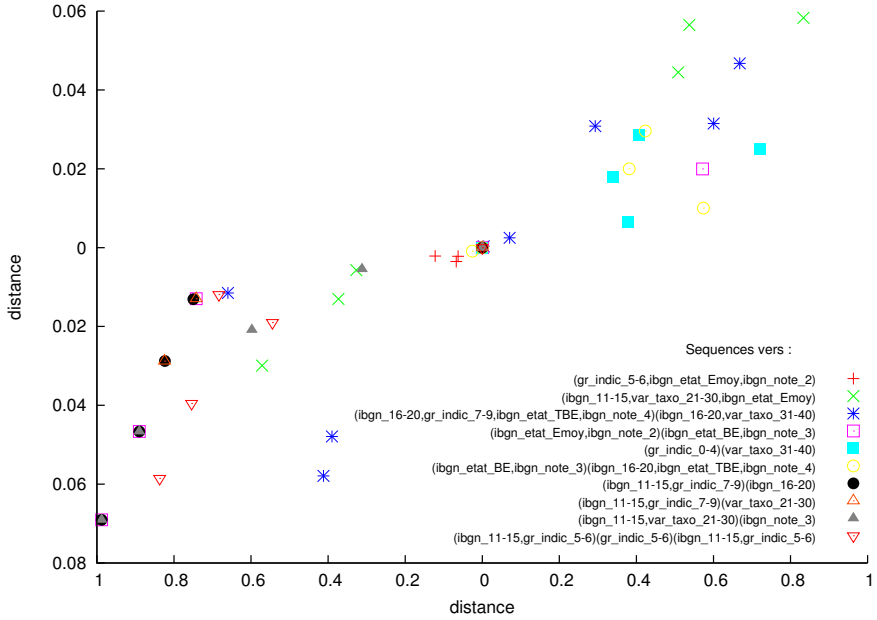

FIG. 8 – *Distance des séquences à la séquence d'intérêt pour le zonage k-voisin*

Les résultats obtenus soulignent la difficulté de fouiller des données sans en connaître réellement tous les contours.

Les perspectives de ce travail sont nombreuses. Tout d'abord, en ce qui concerne les données traitées, des éléments complémentaires relatifs à la détermination des pressions sur l'eau sont en cours d'acquisition. En effet, la détermination exacte de l'état du cours d'eau nécessite de disposer d'autres indicateurs qui ne sont pas présents dans les données actuellement étudiées : l'IPR (Indice Poisson Rivière) et l'IBMR (Indice Biologique Macrophytique en Rivière). Ensuite, pour la phase d'extraction, nous souhaitons comparer différentes techniques de fouille de données en terme de motifs obtenus. En particulier, nous nous focaliserons sur la recherche de motifs spatialisés inspirés des travaux de (Flouvat et al., 2010) comme outil d'analyse et d'exploration des données hydrologiques. Puis nous étendrons cette démarche aux données de pression, caractérisées par l'occupation du sol et des données d'enquêtes. Les questions méthodologiques sont nombreuses : Comment décrire les pressions sur les cours à partir de l'occupation du sol ? Comment modéliser les relations entre occupation du sol et

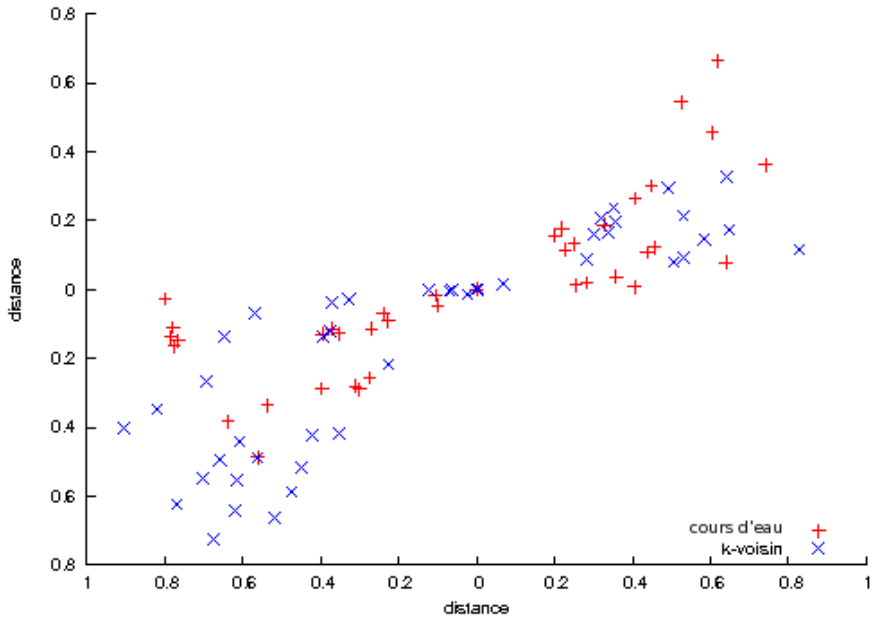

FIG. 9 – *Distance des séquences à la séquence d'intérêt pour les zonages cours d'eau et k-voisin*

qualité des rivières ? Et comment prendre en compte l'hétérogénéité des données ?

Références

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In P. S. Yu et A. L. P. Chen (Eds.), *Proceedings of the Eleventh International Conference on Data Engineering, March 6-10, 1995, Taipei, Taiwan*, pp. 3–14. IEEE Computer Society.

Azé, J. (2003). Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances. In *Revue RIA-ECA numéro spécial EGC03*, Volume 17, pp. 171–182.

Azé, J., P. Lenca, S. Lallich, et BenoîtVaillant (2007). A study of the robustness of association rules. In R. Stahlbock, S. F. Crone, et C. P. S. Lessmann (Eds.), *The 2007 International Conference on Data Mining (DMIN'07)*, pp. 132–137.

Cao, H., N. Mamoulis, et D. W. Cheung (2005). Mining frequent spatio-temporal sequential patterns. In *ICDM'05*, pp. 82–89.

Celik, M., S. Shekhar, J. P. Rogers, J. A. Shine, et J. S. Yoo (2006). Mixed-drove spatio-temporal co-occurence pattern mining : A summary of results. In *ICDM'06*, pp. 119–128.

Flouvat, F., N. Selmaoui-Folcher, et D. Gay (2010). Vers une extraction et une visualisation des co-localisations adaptées aux experts. In *EGC*, pp. 441–452.

Geng, L. et H. J. Hamilton (2006). Interestingness measures for data mining : A survey. *ACM Comput. Surv. 38*.

Gras, R., P. Kuntz, et H. Briand (2001). Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille des données. *Revue Mathématique et Sciences Humaines 154-155*, 9–29.

Jalali-Heravi, M. et O. R. Zaïane (2010). A study on interestingness measures for associative classifiers. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, New York, NY, USA, pp. 1039–1046. ACM.

Julea., A., N. Meger, et P. Bolon (2008). On mining pixel based evolution classes in satellite image time series. In *Proc. of the 5th Conf. on Image Information Mining : pursuing automation of geospatial intelligence for environment and security (ESA-EUSC 2008)*, pp. 6.

Lerman, I.-C. et J. Azé (2007). *A New Probabilistic Measure of Interestingness for Association Rules, Based on the Likelihood of the Link*, pp. 207–236. Springer.

Lerman, I.-C., R. Gras, et H. Rostam (1981). élaboration et évaluation d'un indice d'implication pour des données binaires i. *Revue Mathématique et Sciences Humaines 75*, 5–35.

Lerman, I.-C. et S. Guillaume (2011). Comparaison entre deux indices pour l'évaluation probabiliste discriminante des règles d'association. In RNTI (Ed.), *Conférence Extraction et Gestion des Connaissances (EGC'2011)*, pp. 647–656.

Mabit, L., N. Selmaoui-Folcher, et F. Flouvat (2011). Modélisation de la dynamique de phénomènes spatio-temporels par des séquences. In *EGC*, pp. 455–466.

Marascu, A. et F. Masseglia (2006). Mining sequential patterns from data streams : a centroid approach. *Journal of Intelligent Information Systems 27*(3), 291–307.

Masseglia, F., P. Poncelet, M. Teisseire, et A. Marascu (2008). Web usage mining : extracting unexpected periods from web logs. *Data Mining and Knowledge Discovery (DMKD) 16*(1), 39–65.

Pei, J., J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, et M.-C. Hsu (2004). Mining sequential patterns by pattern-growth : The prefixspan approach. *IEEE Transactions on Knowledge and Data Engineering 16*(11), 2004.

Pei, J., J. Han, B. Mortazavi-Asl, et H. Zhu (2000). Mining access patterns efficiently from web logs. In T. Terano, H. Liu, et A. L. P. Chen (Eds.), *Knowledge Discovery and Data Mining, Current Issues and New Applications, 4th Pacific-Asia Conference, PADKK 2000, Kyoto, Japan, April 18-20, 2000, Proceedings*, Lecture Notes in Computer Science, pp. 396–407. Springer.

Perera, D., J. Kay, I. Koprinska, K. Yacef, et O. R. Zaïane (2009). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering 21*(6), 759–772.

Rabatel, J., S. Bringay, et P. Poncelet (2010). Aide à la décision pour la maintenance ferroviaire préventive. In S. B. Yahia et J.-M. Petit (Eds.), *Extraction et gestion des connaissances (EGC'2010), Actes, 26 au 29 janvier 2010, Hammamet, Tunisie*, Revue des Nouvelles Technologies de l'Information, pp. 363–368. Cépaduès-Éditions.

Salle, P., S. Bringay, et M. Teisseire (2009). Mining discriminant sequential patterns for aging

brain. In C. Combi, Y. Shahar, et A. Abu-Hanna (Eds.), *Artificial Intelligence in Medicine, 12th Conference on Artificial Intelligence in Medicine, AIME 2009, Verona, Italy, July 18-22, 2009. Proceedings*, Lecture Notes in Computer Science, pp. 365–369.

Saneifar, H., S. Bringay, A. Laurent, et M. Teisseire (2008). S2mp : Similarity measure for sequential patterns. In J. F. Roddick, J. Li, P. Christen, et P. J. Kennedy (Eds.), *AusDM*, Volume 87 of *CRPIT*, pp. 95–104. Australian Computer Society.

Tan, P.-N., V. Kumar, et J. Srivastava (2002). Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'02)*, pp. 32–41.

Wang, K., Y. Xu, et J. X. Yu (2004). Scalable sequential pattern mining for biological sequences. In *CIKM '04 : Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management*, New York, NY, USA, pp. 178–187. ACM.

Summary

In this paper, we propose a knowledge discovery process for hydrologic data. We adopt a sequential pattern algorithm applied on data from stations located along several rivers. Data are pre-treated according to different spatial proximities and the analysis of the number of patterns obtained highlights the influence of such defined relations. We propose an objective measure of validation, called *the measure of temporal contradictionless*, to help the expert to discover usual information. With this work, we move towards a new way of spatial indicator discovery to help interpretation of ecological monitoring of watercourses and pressure data.