Exploring Heterogeneous Sequential Data on River Networks with Relational Concept Analysis

Cristina Nica, Agnès Braud, Florence Le Ber

Exploring Heterogeneous Sequential Data on River Networks with Relational Concept Analysis

Cristina Nica, Agnès Braud, Florence Le Ber

ICube, Université de Strasbourg, CNRS, ENGEES
nica.cristina87@gmail.com, agnes.braud@unistra.fr,
florence.leber@engees.unistra.fr
http://icube-sdc.unistra.fr

Abstract. Nowadays, many heterogeneous relational data are stored in databases to be further explored for discovering meaningful patterns. Such databases exist in various domains and we focus here on river monitoring. In this paper, a limited number of river sites that make up a river network (seen as a directed graph) is given. Periodically, for each river site three types of data are collected. Our aim is to reveal user-friendly results for visualising the intrinsic structure of these data. To that end, we present an approach for exploring heterogeneous sequential data using Relational Concept Analysis. The main objective is to enhance the evaluation step by extracting heterogeneous closed partially-ordered patterns organised into a hierarchy. The experiments and qualitative interpretations show that our method outputs instructive results for the hydro-ecological domain.

1 Introduction

In Europe, according to the recomandations of Water Framework Directive [4], a special attention should be given to preserving or restoring the good state of waterbodies. Monitoring and assessing the effect of pollution sources and the one of restoration processes must be done in order to improve domain knowledge and to define guidelines for stakeholders.

During an interdisciplinary research project, namely REX[1], many and various hydro-ecological data have been collected periodically between 2002 and 2014 from a river network (seen as a directed graph). These data are about past restoration projects, temporal evolution of aquatic ecosystems and land use pressures. The REX data have been studied with statistical methods, but relational information could not be taken into account (e.g. effect of upper restoration).

Therefore, in this paper we deal with heterogeneous sequential data and we try to make sense of them by means of hierarchies of heterogeneous closed partially-ordered patterns (cpo-patterns, [2]) that exhibit the natural structure of these data. Indeed, a cpo-pattern is compact, contains the same information

[1] http://obs-rhin.engees.eu

as the set of sequential patterns it synthesises and is user-friendly thanks to its representation as a directed acyclic graph. Moreover, a hierarchy provides a convenient way for navigating to interesting heterogeneous cpo-patterns.

To that end, we extend our self-contained approach RCA-Seq – introduced in [7] and based on Relational Concept Analysis (RCA, [12]) – for exploring classical sequential data to exploring heterogeneous sequential data. We propose to manipulate the data as a directed graph that has heterogeneous itemsets (i.e. a set of itemsets of different domains) as vertices and binary spatial relations as edges. Accordingly, we show that RCA-Seq is robust and can be appropriate for exploring graphs and networks, as well.

The paper is structured as follows. Section 2 gives the theoretical background of our work. Section 3 describes the analysed heterogeneous hydro-ecological data. Section 4 introduces a data model used to encode the data into the RCA input. Section 5 presents the RCA-based exploration step. Section 6 explains our proposal for directly extracting hierarchies of heterogeneous cpo-patterns from the RCA output. In Sect. 7 experimental results are discussed. Section 8 presents related work. Finally, we conclude the paper in Sect. 9.

2 Preliminaries

2.1 Heterogeneous CPO-Patterns

Let $\mathcal{I} = \{I_1, I_2, ..., I_m\}$ be a set of *items*. An *itemset* $IS = (I_{j1}...I_{jk})$, where $I_{j_i} \in \mathcal{I}$, is an unordered subset of $\mathcal{I}$. An itemset IS with k items is referred to as *k-itemset*. Let $\mathcal{IS}$ be the set of all itemsets built from $\mathcal{I}$. A *sequence* S is a non-empty ordered list of itemsets, $S = \langle IS_1 IS_2 ... IS_p \rangle$ where $IS_j \in \mathcal{IS}$. The sequence S is a *subsequence* of another sequence $S' = \langle IS'_1 IS'_2 ... IS'_q \rangle$, denoted as $S \preceq_s S'$, if $p \leq q$ and if there are integers $j_1 < j_2 < ... < j_k < ... < j_p$ such that $IS_1 \subseteq IS'_{j_1}, IS_2 \subseteq IS'_{j_2}, ..., IS_p \subseteq IS'_{j_p}$.

Sequential patterns have been defined by [1] as frequent subsequences discovered in a sequence database. A sequential pattern is associated with a support θ, i.e. the number of sequences containing the pattern.

Suppose now that there is a *partial order* (i.e. a reflexive, antisymmetric and transitive binary relation) on the items, denoted by $(\mathcal{I}, \leq)$. We say that $(\mathcal{I}, \leq)$ is a *poset*. A *multilevel itemset* $IS_{ml} = (I_{j1}...I_{jk})$, where $I_{j_i} \in \mathcal{I}$ and $\nexists I_{j_i}, I_{j_{i'}} \in IS_{ml}$ such that $I_{j_i} \leq I_{j_{i'}}$, is a non-empty and unordered set of items that can be at different levels of granularity (i.e. items from different levels of poset $(\mathcal{I}, \leq)$). We denote by $\mathcal{IS}_{ml}$ the set of all multilevel itemsets built from $(\mathcal{I}, \leq)$. The partial order on the set of all multilevel itemsets $(\mathcal{IS}_{ml}, \subseteq_{ml})$ is defined as follows: $IS_{ml} \subseteq IS'_{ml}$ if $\forall I_j \in IS_{ml}, \exists I_{j'} \in IS'_{ml}, I_{j'} \leq I_j$ and $\forall I_l \neq I_j$, $\exists I_{l'} \neq I_{j'}$ such that $I_{l'} \leq I_l$.

To illustrate this, let us consider $\mathcal{I}_1 = \{a, b, c, d, e, \mathit{Consonants}, \mathit{Vowels}, \mathit{Letters}\}$ a set of items and $(\mathcal{I}_1, \leq)$ a partial order depicted in Fig.1, where an edge represents the binary relation *is-a*, denoted by $\leq$.

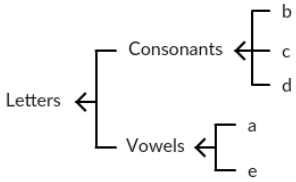


Fig. 1: An example of a partial order on $\mathcal{I}_1 = \{a, b, c, d, e, Consonants, Vowels, Letters\}$

For example, $a \leq Vowels$ designates that letter "a" is a vowel. Let be two itemsets $(a\ b\ c)$ and $(a\ Consonants)$, then $(a\ Consonants) \subseteq_{ml} (a\ b\ c)$ since $a \leq a$ and $b \leq Consonants$ (or $c \leq Consonants$).

Let $\mathcal{H} = \{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_n\}$ be a set of distinct sets of items, where $\mathcal{I}_j$ with $j \in \{1, ..., n\}$ represents a domain. We note that $\mathcal{I}_j$ can be a poset or an unordered set. Let $\mathcal{IS}_j$ be the set of all itemsets built from $\mathcal{I}_j \in \mathcal{H}$. A *heterogeneous itemset* $IS_\mathcal{H} = \{IS_1, IS_2, ..., IS_n\}$, where $IS_j \in \mathcal{IS}_j$, is a non-empty and unordered set of itemsets built from distinct sets of $\mathcal{H}$. In addition, a *multilevel heterogeneous itemset* (hereinafter referred to as heterogeneous itemsets) is a set of itemsets that has at least one multilevel itemset.

Let $\mathcal{IS}_\mathcal{H}$ be the set of all heterogeneous itemsets built from $\mathcal{H}$. The partial order $(\mathcal{IS}_\mathcal{H}, \subseteq_\mathcal{H})$ is defined as follows: $IS_\mathcal{H} \subseteq_\mathcal{H} IS'_\mathcal{H}$ if $\forall IS_k \in IS_\mathcal{H}, \exists IS'_k \in IS'_\mathcal{H}$ such that $IS_k \subseteq IS'_k$, where $IS_k, IS'_k \in \mathcal{IS}_k$, $k \in \{1, ..., n\}$. The order on heterogeneous itemsets is defined accordingly relying on $\subseteq_{ml}$.

To illustrate this, let us consider $\mathcal{H} = \{\mathcal{I}_1, \mathcal{I}_2\}$, where $\mathcal{I}_1$ is partially ordered as shown in Fig. 1 and $\mathcal{I}_2 = \{\square, \lozenge, \triangle\}$ is an unordered set of shapes. Furthermore, let be two multilevel heterogeneous itemsets $IS_{\mathcal{H}_1} = \{(Vowels\ c), (\lozenge)\}$ and $IS_{\mathcal{H}_2} = \{(a\ c), (\square\ \lozenge)\}$, then $IS_{\mathcal{H}_1} \subseteq_\mathcal{H} IS_{\mathcal{H}_2}$ since $(Vowels\ c) \subseteq_{ml} (a\ c)$ (that is $a \leq Vowels$ and $c \leq c$) and $(\lozenge) \subseteq (\square\ \lozenge)$.

A *heterogeneous sequence* $S_\mathcal{H} = \langle IS_{\mathcal{H}_1} IS_{\mathcal{H}_2} ... IS_{\mathcal{H}_r} \rangle$, where $IS_{\mathcal{H}_i} \in \mathcal{IS}_\mathcal{H}$ with $i \in \{1, ..., r\}$, is a non-empty ordered list of heterogeneous itemsets. In addition, a heterogeneous sequence that has at least one multilevel heterogeneous itemset represents a *multilevel heterogeneous sequence* (hereinafter referred to as heterogeneous sequence). A heterogeneous sequence $S_\mathcal{H}$ is a subsequence of another heterogeneous sequence $S'_\mathcal{H} = \langle IS'_{\mathcal{H}_1} IS'_{\mathcal{H}_2} ... IS'_{\mathcal{H}_q} \rangle$, denoted by $S_\mathcal{H} \preceq_{s_\mathcal{H}} S'_\mathcal{H}$, if $r \leq q$ and if there are integers $j_1 < j_2 < ... < j_k < ... < j_r$ such that $IS_{\mathcal{H}_1} \subseteq_\mathcal{H} IS'_{\mathcal{H}_{j_1}}, IS_{\mathcal{H}_2} \subseteq_\mathcal{H} IS'_{\mathcal{H}_{j_2}}, ..., IS_{\mathcal{H}_r} \subseteq_\mathcal{H} IS'_{\mathcal{H}_{j_r}}$. A frequent heterogeneous subsequence is called a *heterogeneous sequential pattern*.

To illustrate this, let be two heterogeneous sequences on the aforementioned $\mathcal{H} = \{\mathcal{I}_1, \mathcal{I}_2\}$: $S1_\mathcal{H} = \langle \{(a\ Consonants), (\square\ \lozenge)\}\ \{(Letters), \emptyset\} \rangle$ and $S2_\mathcal{H} = \langle \{(a\ d), (\square\ \triangle\ \lozenge)\}\ \{(a\ c), (\square)\} \rangle$. Then $S1_\mathcal{H} \preceq_{s_\mathcal{H}} S2_\mathcal{H}$ since

- $\{(a\ Consonants), (\square\ \lozenge)\} \subseteq_\mathcal{H} \{(a\ d), (\square\ \triangle\ \lozenge)\}$, i.e. $a \leq a$, $d \leq Consonants$, $(\square\ \lozenge) \subseteq (\square\ \triangle\ \lozenge)$,
- $\{(Letters), \emptyset\} \subseteq_\mathcal{H} \{(a\ c), (\square)\}$, i.e. $a \leq Letters$ (or $c \leq Letters$), $\emptyset \subseteq (\square)$.

Partially ordered patterns, *po-patterns*, have been introduced by [2], to synthesise sets of sequential patterns. Formally, a *po-pattern* is a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, l)$. $\mathcal{V}$ is the set of vertices, $\mathcal{E}$ is the set of directed edges such that $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, and l is the labelling function mapping each vertex to an itemset. With such a structure, we can determine a strict partial order on vertices u and v such that $u \neq v : u < v$ if there is a directed path from *tail* vertex u to *head* vertex v. However, if there is no directed path from u to v, these elements are not comparable. Each path of the graph represents a sequential pattern, and the set of paths in $\mathcal{G}$ is denoted by $\mathcal{P}_{\mathcal{G}}$. A po-pattern is associated to the set of sequences $\mathcal{S}_{\mathcal{G}}$ that contain all paths of $\mathcal{P}_{\mathcal{G}}$. The support of a po-pattern is defined as $Support(\mathcal{G}) = |\mathcal{S}_{\mathcal{G}}| = |\{S \in \mathcal{D}_{\mathcal{S}} | \forall M \in \mathcal{P}_{\mathcal{G}}, M \preceq_s S\}|$. Furthermore, let $\mathcal{G}$ and $\mathcal{G}'$ be two po-patterns with $\mathcal{P}_{\mathcal{G}}$ and $\mathcal{P}_{\mathcal{G}'}$ their sets of paths. $\mathcal{G}'$ is a sub po-pattern of $\mathcal{G}$, denoted by $\mathcal{G}' \preceq_g \mathcal{G}$, if $\forall M' \in \mathcal{P}_{\mathcal{G}'}, \exists M \in \mathcal{P}_{\mathcal{G}}$ such that $M' \preceq_s M$. A po-pattern $\mathcal{G}$ is *closed*, referred to as *cpo-pattern*, if there exists no po-pattern $\mathcal{G}'$ such that $\mathcal{G} \prec_g \mathcal{G}'$ with $\mathcal{S}_{\mathcal{G}} = \mathcal{S}_{\mathcal{G}'}$. A cpo-pattern whose paths are heterogeneous sequential patterns is called *heterogeneous cpo-pattern*.

2.2 RCA

RCA extends the purpose of Formal Concept Analysis (FCA, [5]) to relational data. RCA applies iteratively FCA on a Relational Context Family (RCF). An RCF is a pair $(\mathcal{K}, \mathcal{R})$, where $\mathcal{K}$ is a set of object-attribute contexts and $\mathcal{R}$ is a set of object-object contexts. $\mathcal{K}$ contains n object-attribute contexts $K_i = (G_i, M_i, I_i), i \in \{1, ..., n\}$. $\mathcal{R}$ contains m object-object contexts $R_j = (G_k, G_l, r_j), j \in \{1, ..., m\}$, where $r_j \subseteq G_k \times G_l$ is a binary relation with $k, l \in \{1, ..., n\}$, $G_k = dom(r_j)$ the domain of the relation, and $G_l = ran(r_j)$ the range of the relation. G_k and G_l are the sets of objects of the object-attribute contexts K_k and K_l, respectively. RCA relies on a relational scaling mechanism that is used to transform a relation r_j into a set of *relational attributes* that extends the object-attribute context describing the set of objects $dom(r_j)$. A relational attribute $\exists r_j(C)$, where $\exists$ is the existential quantifier, and $C = (X, Y)$ is a concept whose extent contains objects from $ran(r_j)$, is owned by an object $g \in dom(r_j)$ if $r_j(g) \cap X \neq \emptyset$. Other quantifiers can be found in [12]. RCA process consists in applying FCA first on each object-attribute context of an RCF, and then iteratively on each object-attribute context extended by the relational attributes created using the learnt concepts from the previous step. The RCA result is obtained when the families of lattices of two consecutive steps are isomorphic and the object-attribute contexts are unchanged.

3 Heterogeneous Hydro-Ecological Data

We focus on hydro-ecological data concerning Rhine river. These data have been collected during REX project. A number of 15 *river sites* (i.e. fixed points) in the Alsace plain were monitored between 2002 – 2014. These sites make up the

Fig. 2: River network

river network illustrated in Fig. 2 that can be seen as a graph of river sites linked by a spatial relation *is downstream of*.

There are three monitored periods of time: 2002 – 2005 (I), 2006 – 2009 (II) and 2010 – 2014 (III). Periodically, for each river site a heterogeneous itemset {*physico-chemical* (PHC) *parameters*, *biological* (BIO) *indicators*, *land use*} is gathered. *PHC parameters* (e.g. temperature, nitrite and dissolved oxygen) indicate the presence or absence of different types of pollutions (e.g. organic or nutrient) according to the qualitative values of parameters. *BIO indicators* (e.g. Standardised Global Biological Index (IBGN), Biological Index of Diatoms (IBD) and Fish Biotic Index (IPR)) determine the quality of water. The indicators and parameters have five qualitative values provided by SEQ-Eau[2] standard, namely *very good*, *good*, *medium*, *bad* and *very bad* represented respectively by the colours *blue*, *green*, *yellow*, *orange* and *red*. All types of *land use* (e.g. forests and urban areas) effect positively or negatively the water quality. The land use around each monitored river site is assessed within two increasing buffers, precisely 100 m and 500 m. A type of land use, e.g. buildings, has a qualitative value according to a percentage of area j covered by it as follows: *low* if $j \in [0\%, 25\%]$, *medium* if $j \in (25\%, 52\%]$ and *high* if $j \in (52\%, 100\%]$. These domains are described by means of taxonomies as shown in Fig. 3. Let us note that the collected data concern only the atomic values from these taxonomies (e.g. urban areas).

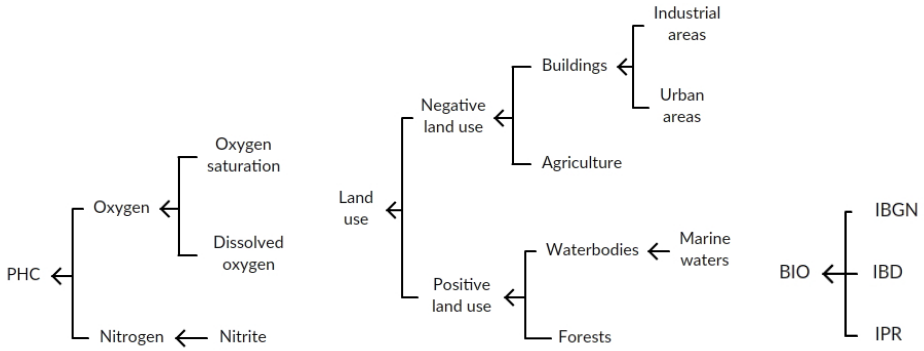
In addition, a river site is included in a *river segment* that can be restored at one or more locations during the whole monitored period of time. There are two types of restoration: *global* and *wetland*. According to the number of restorations i undertaken during 2002 – 2014, there are three *levels of the type of restoration* as follows: $L1$ if $i \in (0, 2]$, $L2$ if $i \in (2, 5]$ and $L3$ if $i \in (5, \infty)$.

For instance, by analysing Fig. 4, the heterogeneous itemset {($\texttt{NITRITE}_{\texttt{red}}$), ($\texttt{IBGN}_{\texttt{green}}$), ($\texttt{FORESTS}_{\texttt{low\_500m}}$ $\texttt{INDUSTRIAL AREAS}_{\texttt{high\_500m}}$)} is associated with river site *S7742* in period 2010 – 2014; the itemset ($\texttt{Wetland}_{\texttt{L1}}$ $\texttt{Global}_{\texttt{L2}}$) is associated with river segment *20165* in the whole monitored period.

4 Data Modelling

Our purpose is to highlight how the ecological state of aquatic ecosystem and the land use in upstream river sites impact the aquatic ecosystem in downstream river sites and, thus determine the necessity of the restorations of river segments.

[2] http://rhin-meuse.eaufrance.fr/IMG/pdf/grilles-seq-eau-v2.pdf



(a) PHC parameters (b) land use (c) BIO indicators

Fig. 3: Excerpts from the taxonomies over the three analysed domains

To that end, by exploiting the relational nature of the various collected hydro-ecological data, we propose the data model shown in Fig. 5. This is used to encode the analysed data into the RCA input. The six rectangles represent the six sets of objects we manipulate: river segments, river sites, restoration types, BIO indicators, PHC parameters and land use. These sets are given in Tab. 1. The set of river sites contains all ordered pairs given by the Cartesian product of the river sites shown in Fig. 2 and the three monitored periods of time. BIO indicators, PHC parameters and land use correspond to the taxonomies depicted in Fig. 3. In addition, let us mention that we consider, firstly, river segments as *target objects* since these are restored; and secondly, river sites as *non-target objects* since these are assessed to understand the necessity of restorations.

Table 1: Analysed sets of objects; the set of river sites is given by the Cartesian product of the river sites given in Fig. 2 and the three monitored periods of time

Set	**Objects**
river sites	{S7778, S8763, S7741, S6702, S7872, S7742, S7914, S7915, S7743, S7792, S9515, S8027, S7753, S7965, S8735} × {I, II, III}
river segments	3163, 4548, 5601, 6850, 8614, 8674, 18725, 19754, 19949, 20165, 20346, 26763
land use	LAND USE, NEGATIVE LAND USE, POSITIVE LAND USE, BUILDINGS, AGRICULTURE, URBAN AREAS, INDUSTRIAL AREAS, LANDFILL & MINE SITES, ARABLE LANDS, PERMANENT CROPS, FORESTS & NATURAL AREAS, FORESTS, HERBACEOUS PLANTS, WETLANDS, WATERBODIES, CONTINENTAL WATERS, MARINE WATERS
PHC parameters	PHC, NITROGEN, NITRITE, AMMONIUM, PHOS, TOTAL PHOSPHORUS, NITRATE, OXYGEN, OXYGEN SATURATION, DISSOLVED OXYGEN, BIOLOGICAL OXYGEN DEMAND, TEMPERATURE
BIO indicators	IBGN, IBD, IPR
restoration types	Wetland, Global

The links between objects are highlighted by using binary relations as follows:

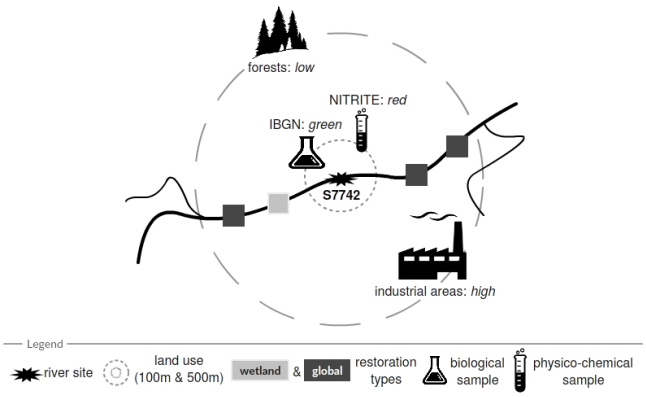


Fig. 4: River segment *20165*; period III: river site *S7742*

- spatial relation *includes* associates a river site with a river segment if the river site is in the river segment;
- spatial relation *is downstream of* is used to encode into the RCA input the river network shown in Fig. 2;
- qualitative relation *has restoration L1/L2/L3* associates a river segment with the type of undertaken restoration;
- qualitative relation *has indicator blue/green/yellow/orange/red* associates a river site with a measured BIO indicator;
- qualitative relation *has parameter blue/green/yellow/orange/red* associates a river site with a measured PHC parameter;
- spatial-qualitative relation *is surrounded by low_100m/low_500m/medium_100m/medium_500m/high_100m/high_500m* associates a river site with a type of land use.

5 Exploration of Heterogeneous Data by Using RCA

In this section we briefly recall and slightly adapt the RCA-exploration step of sequential data presented in our previous paper [7].

Firstly, the RCA input (RCF) – an excerpt is depicted in Tab. 2 – is built by relying on the data model shown in Fig. 5 and on the sets of objects given in Tab. 1. Basically, this RCF encodes all hydro-ecological data collected during the whole monitored period 2002 – 2014. There is an object-attribute context for each rectangle out of the data model, precisely `KSegments` (river segments), `KSites` (river sites), `KRT` (restoration types), `KBIO` (BIO indicators), `KPHC` (PHC parameters) and `KLU` (land use). `KSites` has no column since river sites are described only by using the *has indicator*, *has parameter* and *is surrounded by* relations. Similarly, `KSegments` has no column since river segments are described by using the *has restoration* relations. As shown in Tab. 2, a nominal scaling is

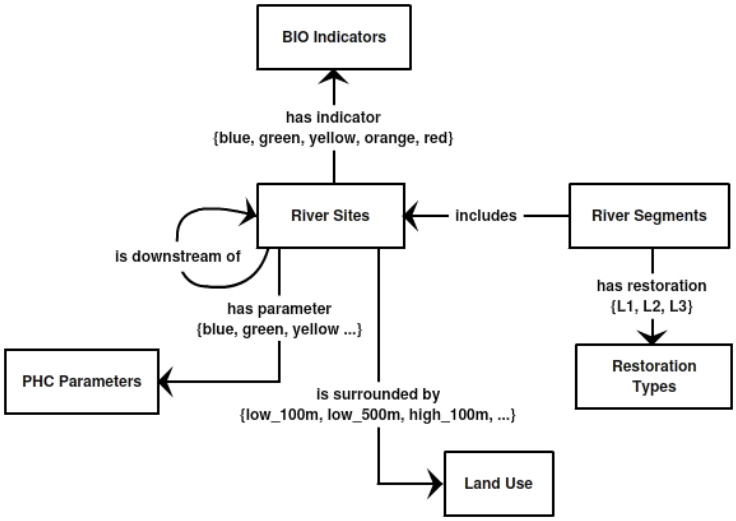


Fig. 5: Modelling heterogeneous hydro-ecological data

used to build `KRT` in order to obtain a partial order over the unordered set of restoration types. In contrast, an ordinal scaling is used to build `KBIO`, `KPHC` and `KLU` in order to encode the taxonomies given in Fig. 3. In addition, there are 21 object-object contexts, one for each relation out of the data model, e.g. in Tab. 2 `RSite-ds-Site` (river site *is downstream of* river site) and `RSite-red-BIO` (river site *has red* BIO *indicator*).

Secondly, RCA is applied[3] to the RCF shown in Tab. 2 and a family of concept lattices is obtained after four iterations. The RCA output comprises six concept lattices, one for each object-attribute context, as follows: *target lattice* $\mathcal{L}_{\texttt{KSegments}}$ (river segments), *non-target lattice* $\mathcal{L}_{\texttt{KSites}}$ (river sites), *lattice of restoration types* $\mathcal{L}_{\texttt{KRT}}$ and the *taxonomy lattices* $\mathcal{L}_{\texttt{KBIO}}$, $\mathcal{L}_{\texttt{KPHC}}$, $\mathcal{L}_{\texttt{KLU}}$ that correspond to the taxonomies illustrated in Fig. 3. The concepts of the latter three lattices are used to describe river sites by means of the revealed qualitative relational attributes. Similarly, the concepts of $\mathcal{L}_{\texttt{KRT}}$ are used to describe river segments.

It is worthwhile to mention that the RCA-based exploration step employs a relational scaling mechanism that relies on quantifier $\exists$ because the objective is to capture all the relations between the analysed objects. The target lattice and non-target one contain respectively 860 and 4554 concepts.

[3] using `http://dolques.free.fr/rcaexplore`

Table 2: Excerpt of the RCA input composed of object-attribute contexts (KSites, KRT and KPHC) and object-object contexts (RSite-ds-Site and RSite-red-BIO).

KSites
(S7743,I)
(S7743,II)
(S7743,III)
(S8735,I))
(S8735,II)

KRT	Wetland	Global
Wetland	×	
Global		×

KPHC	PHC	NITROGEN	NITRITE	PHOS
PHC	×			
NITROGEN	×	×		
NITRITE	×	×	×	
PHOS	×			×

RSite-ds-Site	(S7872,I)	(S7872,II)	(S9515,I)
(S7743,I)	×		
(S7743,II)		×	
(S7743,III)			
(S8735,I))	×		×
(S8735,II)		×	

RSite-red-PHC	PHC	NITROGEN	NITRITE
(S7743,I)			
(S7743,II)			×
(S7743,III)			
(S8735,I))			
(S8735,II)			×

6 Extraction of Heterogeneous CPO-Patterns Organised into a Hierarchy

To extract a hierarchy of heterogeneous cpo-patterns from the RCA output (obtained as explained in Sect. 5), we apply and slightly modify the RCA-Seq approach presented in [7]. Let us note that the hierarchy is directly obtained since each concept of the target lattice is associated with a heterogeneous cpo-pattern.

Briefly, starting with a concept from the target lattice, a heterogeneous cpo-pattern is extracted by navigating interrelated concept intents. For each navigated concept intent, a vertex (labelled with an itemset) is derived from all (spatial-)qualitative relational attributes (hereinafter referred to as qualitative relational attributes) whereas an edge is derived from a spatial relational attribute. A qualitative/spatial relational attribute highlights a qualitative/spatial relation.

In fact, in this paper a vertex derived from a concept intent of the non-target lattice (river sites) is actually a *heterogeneous vertex* labelled with a heterogeneous itemset. Basically, an itemset of the heterogeneous itemset is built for each set of qualitative relational attributes, which define the same qualitative relation, out of the concept intent. Therefore, for a concept intent we analyse the qualitative relational attributes, which are built using a qualitative relation r_q and concepts from a *taxonomy lattice* $\mathcal{L}_{K_{tax}} = (\mathcal{C}_{K_{tax}}, \preceq_{K_{tax}})$, to derive items as follows:

- from a qualitative relational attribute $\exists r_q(C_{tax})$, where $C_{tax} \in \mathcal{C}_{K_{tax}}$, is derived an item, denoted by "$item_q$", where $extent(C_{tax}) = \{item\}$ and q is the item quality according to r_q;
- if there is no qualitative relational attribute that highlights relation r_q and the information introduced by this relation is mandatory, then is derived an item, denoted by "$item_?$" where $extent(\top(\mathcal{L}_{K_{tax}})) = \{item\}$, that constitutes the 1-itemset obtained for this type of information; conversely, if the

information introduced by this relation is not mandatory, then no item is derived and, thus $\emptyset$ is obtained for this type of information.

Let us mention that a vertex derived from a concept intent of the target lattice (river segments) is labelled with a multilevel itemset. As described in [7], this itemset can contain: the abstract item ("$?_{?}$" – different types of restoration at distinct number of locations), qualitative abstract items (e.g. "$?_{\texttt{L1}}$" – different types of restoration at most 2 locations) and/or concrete items (e.g. "$\texttt{Global}_{\texttt{L3}}$" – global restorations at more than 5 locations).

7 Results and Evaluation

In this section, we present some interesting results obtained with the RCA-Seq approach applied to the heterogeneous data collected between 2002 – 2014 from the river network shown in Fig. 2. The evaluation relies on the positive feedback given by a hydro-ecologist who is well acquainted with cpo-patterns.

By navigating the lattices starting from the target concepts of $\mathcal{L}_{\texttt{KSegments}}$ we obtain a hierarchy of 859 heterogeneous cpo-patterns (the bottom concept of $\mathcal{L}_{\texttt{KSegments}}$ is not considered since generally it is too specific and associated with no river segment). It is worthwhile to mention that a smaller hierarchy of cpo-patterns can be extracted by varying the quantifiers employed by the relational scaling mechanism. In addition, various measures [6] can be used to select relevant heterogeneous cpo-patterns.

Figure 6 depicts an excerpt from this hierarchy, precisely the organised ①, ②, ③, ④, ⑤, ⑥ and ⑦ heterogeneous cpo-patterns.

A cpo-pattern is associated with a set of river segments whose number (support) is shown in ■. The restoration types of these river segments are illustrated in ▭ , e.g. $\texttt{Global}_{\texttt{L1}}$ meaning that the river segments have at most 2 locations with global restoration. A vertex (⬭) is associated with a set of river sites and it is labelled with PHC parameters and their qualitative values. A vertex can have additional information: land use (⬭) and BIO indicators (◊). In the following, we focus on the cpo-patterns ①, ④ and ⑥.

CPO-Pattern ① is associated with 11 (■ in Fig. 6) river segments that contain at most 2 locations with global restoration. In addition, itemset ($\texttt{PHC}_{\texttt{blue}}$) reveals locally (i.e. in the associated river segments) a very good PHC state of water.

CPO-Pattern ④ is associated with 5 river segments that contain at most 2 locations with global restoration. Itemset ($\texttt{IBD}_{\texttt{green}}$) (◊ in Fig. 6) reveals locally a good ecological state of water based on the analysis of diatom species. In addition, the PHC state of water is very good for temperature, biological oxygen demand and nitrogen that represent a part of the abiotic characteristics suitable for diatom species [11].

CPO-Pattern ⑥, which is a more concrete specialisation of ⑤, is associated with 3 river segments that contain at most 2 locations with global and wetland restorations. Itemset ($\texttt{BIO}_{\texttt{green}}$) (◊ in Fig. 6) reveals locally a good ecological state of the aquatic ecosystem. Since BIO is an abstract item, we cannot

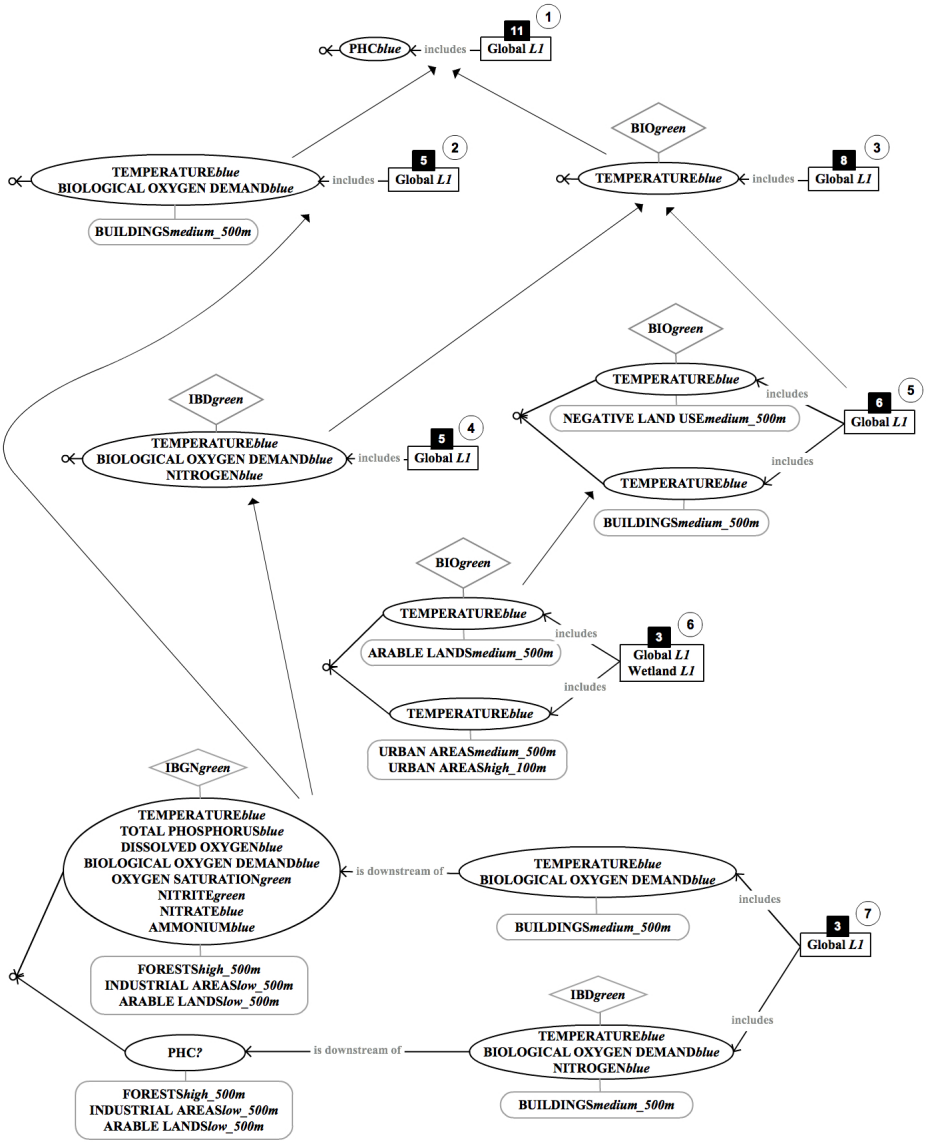


Fig. 6: Excerpt from the hierarchy of heterogeneous cpo-patterns discovered in the analysed hydro-ecological data. ①, ②, ③, ④, ⑤, ⑥ and ⑦ identify the cpo-patterns. ■ is the support (number of river segments) of a cpo-pattern; ▭ represents the types of river segment restoration; ⬭ represents land use; ◊ represents BIO indicators; ⬭ represents PHC parameters

specify the fauna and flora that underpin this regularity. In addition, itemset ($\texttt{TEMPERATURE}_{\texttt{blue}}$) reveals locally a very good PHC state of the water temperature. Furthermore, locally at 500 m buffer the land use pressures of arable lands and urban areas are *medium* whereas at 100 m the land use pressures of urban areas are *high*.

Figure 7 depicts a more complex heterogeneous cpo-pattern. This is associated with the river segments *8674* and *19949* that contain at most 2 locations with global restoration.

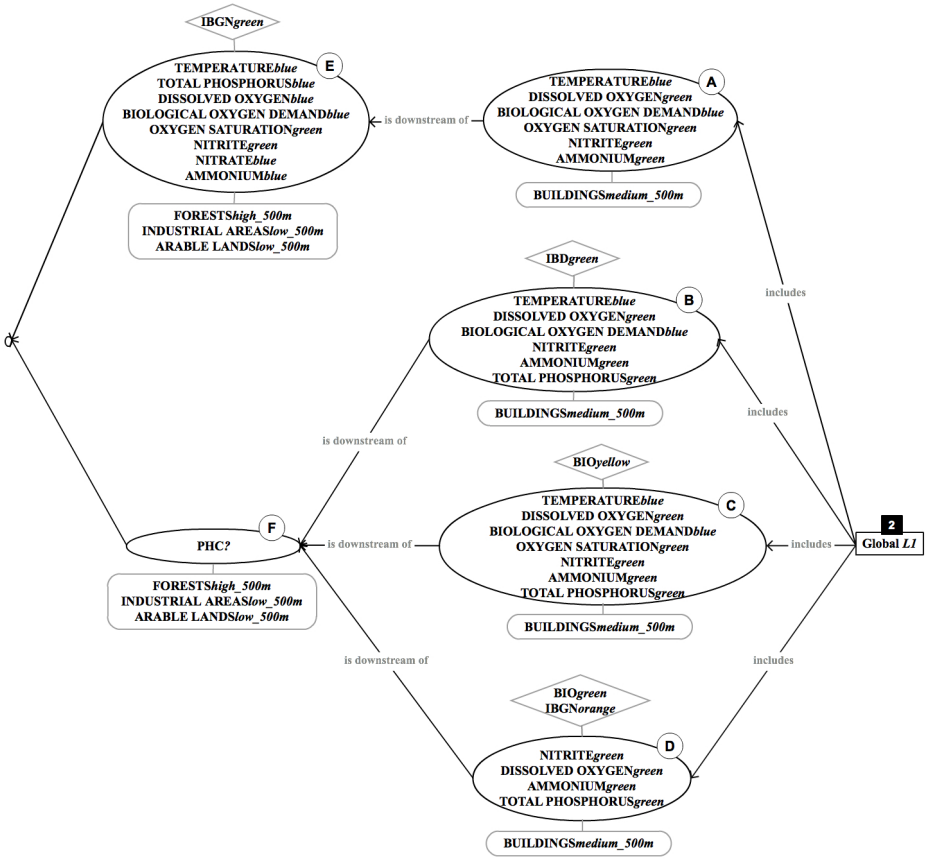


Fig. 7: A complex heterogeneous cpo-pattern extracted from the analysed hydro-ecological data. Ⓐ, Ⓑ, Ⓒ, Ⓓ, Ⓔ and Ⓕ identify the vertices; ■ is the support (number of river segments) of the cpo-pattern; ▭ represents the types of river segment restoration; ⬭ represents land use; ◊ represents BIO indicators; ⬭ represents PHC parameters

The vertices are derived from concepts of $\mathcal{L}_{\texttt{KSites}}$ whose extents are as follows: Ⓐ: {(*S7743*,III), (*S7915*, III)}, Ⓑ: {(*S7915*,II), (*S7743*,III)}, Ⓒ: {(*S7915*, I), (*S7743*,III)}, Ⓓ: {(*S7915*,II), (*S7743*,II)}, Ⓔ: {(*S7914*,III)} and Ⓕ: {(*S7914*,I), (*S7914*,II), (*S7914*,III)}. Locally, in the whole monitored period 2002 – 2014 the land use pressures of buildings are *medium* at 500 m buffer. In contrast, in the upstream rivers at 500 m buffer on the one hand the land use pressures of industrial areas and arable lands are *low*; on the other hand, a *high* percentage of the area is covered with forests that lead to a good ecological state of the aquatic ecosystem in the surroundings. Indeed, by analysing the Ⓔ vertex, itemset ($\texttt{IBGN}_{\texttt{green}}$) (◊, Fig. 7) reveals a good ecological state of the aquatic ecosystem in the period 2010 – 2014 based on the analysis of macro-invertebrates. Moreover, water temperature is very good; organic matter (dissolved oxygen, biological oxygen demand and oxygen saturation) is good and very good; nitrogenous parameters (nitrite and ammonium), which are related to organic matter, are as well good and very good; and nutrients (total phosphorous and nitrate) are very good.

By comparing the Ⓔ vertex with the Ⓐ, Ⓑ, Ⓒ and Ⓓ vertices, it is noted a degradation up to one level regarding the qualitative values of PHC parameters probably caused by the *medium* building pressures at 500 m buffer, e.g.:

- $\texttt{AMMONIUM}_{\texttt{blue}}$ and $\texttt{DISSOLVED OXYGEN}_{\texttt{blue}}$ (very good) from Ⓔ are measured when the surroundings are covered with a low percentage of industrial areas and arable lands (i.e. the land use pressures are low), while $\texttt{AMONIUM}_{\texttt{green}}$ and $\texttt{DISSOLVED OXYGEN}_{\texttt{green}}$ (good) from Ⓐ, Ⓑ, Ⓒ and Ⓓ are measured when the surroundings are covered with a medium percentage of buildings (i.e. the land use pressures are medium);
- $\texttt{TOTAL PHOSPHORUS}_{\texttt{blue}}$ (very good) from Ⓔ is measured when in the surroundings the land use pressures are low; $\texttt{TOTAL PHOSPHORUS}_{\texttt{green}}$ (good) from Ⓑ, Ⓒ and Ⓓ is measured when in the surroundings the land use pressures are medium.

Furthermore, the cpo-pattern shown in Fig. 7 reflects that BIO indicators seem to be more sensitive (up to two levels of their qualitative values) to land use pressures [14,15]. For instance, $\texttt{IBGN}_{\texttt{green}}$ in upstream rivers (Ⓔ) in contrast to $\texttt{BIO}_{\texttt{yellow}}$ and $\texttt{IBGN}_{\texttt{orange}}$ locally (Ⓒ and Ⓓ, respectively).

To sum up, by means of cpo-patterns, we can help hydro-ecologists to check well-known correspondences among the analysed ecological factors as well as to consider lesser-known facts.

8 Related Work

Classical sequential pattern mining approaches deal with sequences whose items are homogeneous and, therefore cannot be applied to heterogeneous sequences (i.e. sequences whose items are different in nature). To our knowledge, [9] proposed the first work for exploring multidimensional sequential data. A multidimensional sequence takes the form $(d_1, d_2, ..., d_m, S)$, where S is a sequence of

itemsets and d_i represents the i^{th} type of information associated with S. The authors proposed three methods for extracting multidimensional sequential patterns that rely on classical sequential pattern algorithms (e.g. PrefixSpan [8]). A key drawback of such multidimensional sequences is the additional information that is constant for all itemsets of sequence S.

In [10], a multidimensional sequence is defined as an ordered list of multidimensional items. A multidimensional item takes the form $(d_1, d_2, ..., d_n)$, where d_k is an item of the k^{th} dimension. Furthermore, each considered dimension is represented at different levels of granularity by means of partial orders. Therefore, multilevel sequential patterns can be discovered, as explained in [13]. The authors proposed the M3SP algorithm that searches for multidimensional and multilevel sequential patterns in two steps. First, the most specific frequent multidimensional items, referred to as *maf-sequences*, are found. Second, the maf-sequences are used to remodel the original multidimensional sequences and then these sequences are mined by using algorithm SPADE [16].

Nevertheless, [3] highlighted a limitation of M3SP, i.e. the multidimensional items do not allow itemsets whose items are of k^{th} dimension. The authors proposed the MMISP algorithm that considers complex and heterogeneous sequences, where a sequence contains *elementary sequences* (ES), i.e. itemsets whose items can be of two types: atomic and different in nature taken from user-defined taxonomies or subsets of unordered sets of items. MMISP does not discover directly sequential patterns in heterogeneous data since a preprocessing step is involved, i.e. the original sequences are encoded into classical sequences. In contrast, RCA-Seq directly searches for cpo-patterns (rather than sequential patterns) in complex and heterogeneous data and, besides, reveals how these patterns relate to each other. Moreover, our approach generalises the ES proposed in [3] by considering its atomic items as 1-itemsets.

9 Conclusion

RCA-Seq is an approach for exploring classical sequential data. In this paper, we have presented an extension of RCA-Seq that highlights its generality, i.e. the capability to explore sequential data regardless of their complexity. Given heterogeneous sequential data on river networks, we have shown that hydro-ecologists can draw valuable insights by exploiting the "richness" (e.g. the additional information captured by concept extents and the revealed abstract items) of the RCA-Seq output. In the future, we plan to (i) improve our extension in order to be applicable to large volumes of heterogeneous sequential data and (ii) explore with RCA-Seq complex relational data such as social networks and knowledge graphs.

Acknowledgement

The REX data were provided by Prof. Jean-Nicolas Beisel and the interpretation of cpo-patterns was done with the help of Corinne Grac (ENGEES Strasbourg).

References

1. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Int. Conference on Data Engineering. pp. 3–14 (1995)
2. Casas-Garriga, G.: Summarizing sequential data with closed partial orders. In: 2005 SIAM Int. Conference on Data Mining. pp. 380–391 (2005)
3. Egho, E., Jay, N., Raïssi, C., Ienco, D., Poncelet, P., Teisseire, M., Napoli, A.: A contribution to the discovery of multidimensional patterns in healthcare trajectories. Journal of Intelligent Information Systems 42(2), 283–305 (2014)
4. European Union: Directive 2000/60/ec of the European parliament and of the council of 23 october 2000 establishing a framework for community action in the field of water policy. Official Journal OJ L 327, 1–73 (2000)
5. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer (1999)
6. Nica, C., Braud, A., Dolques, X., Huchard, M., Ber, F.L.: Exploring temporal data using relational concept analysis: An application to hydroecological data. In: Proceedings of the 13th Int. Conf. on Concept Lattices and Their Applications, CLA 2016. pp. 299–311. CEUR-WS.org (2016)
7. Nica, C., Braud, A., Dolques, X., Huchard, M., Le Ber, F.: Extracting hierarchies of closed partially-ordered patterns using relational concept analysis. In: Graph-Based Representation and Reasoning: 22nd Int. Conf. on Conceptual Structures, ICCS 2016, Proceedings. pp. 17–30. Springer (2016)
8. Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M.C.: Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In: Proceedings of the 17th Int. Conf. on Data Engineering. pp. 215–224. ICDE'01, IEEE Computer Society (2001)
9. Pinto, H., Han, J., Pei, J., Wang, K., Chen, Q., Dayal, U.: Multi-dimensional sequential pattern mining. In: Proceedings of the 10th Int. Conf. on Information and Knowledge Management. pp. 81–88. CIKM, ACM (2001)
10. Plantevit, M., Laurent, A., Laurent, D., Teisseire, M., Choong, Y.W.: Mining multidimensional and multilevel sequential patterns. ACM Trans. Knowl. Discov. Data 4(1), 4:1–4:37 (Jan 2010)
11. Raibole M, S.Y.: Impact of physico-chemical parameters on microbial diversity: Seasonal study. Curr World Environ 6(1), 71–76 (2011)
12. Rouane-Hacene, M., Huchard, M., Napoli, A., Valtchev, P.: Relational concept analysis: Mining concept lattices from multi-relational data. Annals of Mathematics and Artificial Intelligence 67(1), 81–108 (2013)
13. Srikant, R., Agrawal, R.: Mining sequential patterns: Generalizations and performance improvements. In: Proceedings of the 5th Int. Conf. on Extending DB Technology: Advances in DB Technology. pp. 3–17. EDBT, Springer-Verlag (1996)
14. Villeneuve, B., Souchon, Y., Usseglio-Polatera, P., Ferrol, M., Valette, L.: Can we predict biological condition of stream ecosystems? a multi-stressors approach linking three biological indices to physico-chemistry, hydromorphology and land use. Ecological Indicators 48, 88–98 (2015)
15. Wasson, J., Villeneuve, B., Mengin, N., Pella, H., Chandesris, A.: Quelle limite de " bon état écologique " pour les invertébrés benthiques en rivières ? Apport des modèles d'extrapolation spatiale reliant l'indice biologique global normalisé à l'occupation du sol. Ingénieries - E A T 1(47), 3–15 (2006)
16. Zaki, M.J.: Spade: An efficient algorithm for mining frequent sequences. Machine Learning 42(1), 31–60 (2001)