Exploring Temporal Data Using Relational Concept Analysis: An Application to Hydroecology

Cristina Nica[1], Agnès Braud[1], Xavier Dolques[1]
Marianne Huchard[2], Florence Le Ber[1]

[1] ICube, University of Strasbourg, CNRS, ENGEES
firstname.lastname@engees.unistra.fr,agnes.braud@unistra.fr
http://icube-sdc.unistra.fr
[2] LIRMM, University of Montpellier, CNRS
huchard@lirmm.fr
https://www.lirmm.fr

Abstract. This paper presents an approach for mining temporal data, based on Relational Concept Analysis (RCA), that has been developed for a real world application. Our data are sequential samples of biological and physico-chemical parameters taken from watercourses. Our aim is to reveal meaningful relations between the two types of parameters. To this end, we propose a comprehensive temporal data mining process starting by using RCA on an ad hoc temporal data model. The results of RCA are converted into closed partially ordered patterns to provide experts with a synthetic representation of the information contained in the lattice family. Patterns can also be filtered with various measures, exploiting the notion of temporal objects. The process is assessed through some quantitative statistics and qualitative interpretations resulting from experiments carried out on hydroecological datasets.

1 Introduction

Exploring temporal datasets is a major challenge in current research and various methods have therefore been proposed since the 90's [1]. It is worth pointing out that temporal data are relational, so that relational methods [6] can be useful to respect their relational structure, e.g. [9]. In particular, Relational Concept Analysis (RCA, [16]) allows to classify relational data and provides hierarchical results which facilitates the analysis step.

Based on these properties, we propose to use RCA for exploring sequential datasets from the hydroecological domain. These datasets were collected during the Fresqueau project[3] that focused on methods for assessing the quality of watercourses. The collected data represent biological (Bio) and physico-chemical (PhC) samples taken at fixed points (river sites) and repeated in time. Both parameters are used by the experts to determine the quality of watercourses.

[3] http://engees-fresqueau.unistra.fr/presentation.php?lang=en

Therefore, a global assessment of the temporal relationship between PhC and Bio parameters is needed. To this end, preprocessings of the raw sequential data allow to build a qualitative temporal model that can be used to apply RCA on these data. The RCA result is a family of lattices that can be navigated by the users. The users can select relevant navigation paths through the lattices (starting from concepts in a main lattice) by applying measures of interest based on the concept extents, that can be linked to geographical information in our application. Furthermore, in order to help their analysis and to synthetize the results, we propose to transform those concepts within closed partially ordered patterns (cpo-patterns, [5]), i.e. directed acyclic graphs where vertices are labelled with information extracted from the concepts out of the family of lattices. Since concepts can be more or less general or specific, the extracted patterns can be classified within three types, according to the number of vertices that are labelled with general information. Then the users can choose to select and to navigate general or specific paths in the lattices.

The paper is structured as follows. Section 2 presents basic definitions and related work. Section 3 describes the hydroecological data and their preprocessing while the RCA process is detailed in Section 4. Section 5 introduces some measures of interest dealing with the temporal dimension of obtained concepts. Section 6 presents cpo-patterns in order to help the analysis. Section 7 describes and discusses the experimental results carried out on Fresqueau datasets. Section 8 concludes and gives a few perspectives of this work.

2 Basics and Related Work

Relational Concept Analysis (RCA, [16]) extends Formal Concept Analysis (FCA [11]) to classify sets of objects described by attributes and relations, thus allowing to discover knowledge patterns and implication rules in relational datasets. RCA applies iteratively FCA on a Relational Context Family (RCF) that is constituted of a set $\mathcal{K}$ of object-attribute contexts and a set $\mathcal{R}$ of object-object contexts. $\mathcal{K}$ contains n object-attribute formal contexts $K_i = (G_i, M_i, I_i), i \in \{1, ..., n\}$. $\mathcal{R}$ contains m object-object relational contexts $R_j = (G_k, G_l, r_j), j \in \{1, ..., m\}$, where G_k, called the domain of the relation, and G_l, called the range of the relation, are respectively the sets of objects of K_k and K_l, and $r_j \subseteq G_k \times G_l, k, l \in \{1, ..., n\}$. At each step, object-attribute contexts are extended with relational attributes taking the syntactic form $qr_j(C)$, where q is a quantifier, r_j is a relation and $C = (X, Y)$ is a concept where X is a subset of objects from the range of r_j. This paper uses the *existential* quantifier: $\exists r_j(C)$ is an attribute of $o \in G_k$ if $r_j(o) \cap X \neq \emptyset$. 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 concepts from the previous step. The RCA result is obtained when the family of lattices of two consecutive steps are isomorphic and the contexts are unchanged.

RCA has been applied to various data, e.g. for software model analysis and re-engineering [2]. To our knowledge, this is the first time that RCA is used to explore sequential datasets. There are, however, various related FCA approaches. [18] introduced Temporal Concept Analysis where objects are characterized with a date and a state (i.e. a set of attributes). Data are merged into a single context, and the resulting concept lattice is analysed thanks to the date element in the concepts, so that temporal relations between concepts are actually revealed by the analyst. This approach has been used to analyse sequential data about crime suspects [15]. In our RCA approach, the temporal relation between dates is considered as an object-object relation and it links concepts from several lattices. In [8], sequential datasets are processed without involving any partial order. In [5], closed subsequences are mined and then grouped in a lattice similar to a concept lattice. In [4], sequential data are mapped onto pattern structures whose projections are used to build a pattern concept lattice. The authors combine the stability of concepts and the projections of pattern structures in order to select relevant patterns.

Besides, there exist various methods to explore qualitative sequential data. Indeed, sequential pattern mining is an active research area, in relation to the exponential growth of temporal and spatio-temporal databases. Sequential patterns have been introduced by [1] and used for different purposes. Such an approach has been developed within the Fresqueau project and focused on closed po-patterns, which were selected through various measures [7]. Indeed, selecting relevant results is a main challenge for all approaches dealing with large datasets. In FCA, the most used measures for selecting relevant concepts are stability [13], probability and separation [12]. Unfortunately, these measures are not able to take into account the specific structure of concepts built on temporal objects. We thus propose to use specific measures, as detailed in Section 5.

3 Context and Data Preprocessing

In the Fresqueau project, the analysed data cover various compartments such as physico-chemistry, hydrobiology, hydromorphology and land use (as described in [3]). Here, we try to tackle the following issue by means of RCA: *Can experts explain values of biological parameters from PhC values occuring in past months and thus improve the global assessment of the quality of watercourse ecosystems?*

To answer this question we should mention that the quality of watercourses is determined by the Bio parameters (e.g. Standardised Global Biological Index (IBGN), Biological Index of Diatoms (IBD) and Fish Biotic Index (IPR)). Hence, the objects of interest from our work are the Bio samples and we want to assess, over a period of time, the impact of PhC macro-parameters (e.g. Nitrogen (AZOT), Phosphor (PHOS) and Particulate Matter (PAES)) on Bio ones.

Table 1(a) illustrates a small raw sequential dataset of Bio and PhC samples taken from a *site* (e.g. S1) corresponding to a river segment. A set of sites constitutes a *geographical area.* A data sequence is a chronologically ordered set of PhC samples with a Bio one at the end, all taken from the same site. This raw sequential dataset shows measurements made only for IBGN Bio parameter and for four PhC parameters namely Ammonium (NH_4^+), Kjeldahl Nitrogen

Table 1: Small example of raw and corresponding preprocessed sequential dataset.

(a) Raw Sequential Data

Site	**Date**	NH_4^+	NKJ	NO_2^-	PO_4^{3-}	**IBGN**
S1	08/05	-	-	-	-	10
	06/05	0.004	-	0.012	0.035	-
	09/04	-	-	-	-	8
	08/04	-	1.414	-	-	-
	01/04	0.043	0.146	0.421	-	-

(b) Preprocessed Sequential Data

Site	**Date**	**AZOT**	**PHOS**	**IBGN**
S1	08/05	-	-	Yellow
	06/05	Blue	Green	-
	09/04	-	-	Orange
	08/04	Green	-	-

(NKJ), Nitrite (NO_2^-) and Orthophosphate (PO_4^{3-}). For instance, 0.043 mg/l of NH_4^+ is measured on 01/04, i.e. January 2004, for the site $S1$. An IBGN score of 8/20 is measured on September 2004 for the same site.

The raw sequential dataset contains only numerical values. For mining such data, we transform them by applying discretization and selection processes based on domain knowledge. The discretization aims at converting numerical values into qualitative ones. To this end, we use qualitative values for Bio and PhC parameters that are provided by the SEQ-Eau[4] standard. Both types of parameters have five qualitative values, namely *very good*, *good*, *medium*, *bad* and *very bad* represented respectively by the colors *blue*, *green*, *yellow*, *orange* and *red*. In addition, SEQ-Eau standard groups PhC parameters into macro-parameters. For example, NH_4^+, NKJ and NO_2^- are grouped into AZOT macro-parameter. The selection process considers only relevant data by defining some constraints based on expert advice. For instance, the only analysed PhC samples are those taken within *4 months* before a Bio parameter, from the same site.

Table 1(b) shows the preprocessed sequential dataset ready to be mined using RCA. This sequential dataset is obtained by applying the discretization and selection processes to the raw sequential dataset illustrated in Tab. 1(a). It is worth pointing out that the preprocessed sequential dataset is significantly small compared to the raw one thanks to the macro-parameters and the limited analysed period of time.

4 Temporal Relational Analysis

The sequential dataset is structured following the schema depicted in Fig. 1. The four rectangles represent the four sets of objects we manipulate: Bio samples, PhC samples, Bio parameters and PhC parameters. The links between Bio/PhC samples and PhC samples are defined by the temporal binary relation *is preceded by* (denoted by ipb). This temporal relation associates one sample to another one if the first sample is preceded in time by the second one, on the same site. There

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

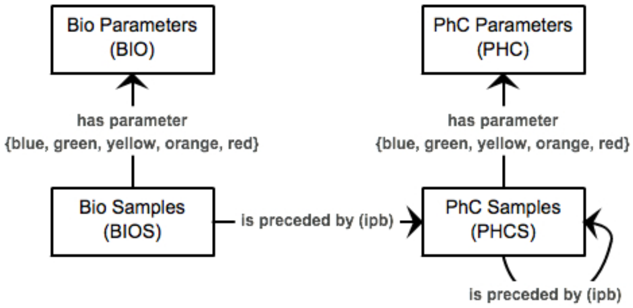


Fig. 1: The modelling of the hydroecological sequential dataset.

is no temporal binary relation between Bio samples since in this work we evaluate the impact of physico-chemistry on biology. The Bio/PhC samples are described only by the qualitative relations *has parameter blue/green/yellow/orange/red* that link the Bio/PhC samples with the measured Bio/PhC parameters. For instance, *has parameter green* links the PhC samples taken from S1 on 08/04 (Tab. 1(b)) with AZOT PhC parameter.

Following the temporal data model illustrated in Fig. 1, we build the RCF depicted in Tab. 2 for a small hydroecological sequential dataset. The tables `KPHC` (PhC parameters), `KBIOS` (Bio samples) and `KPHCS` (PhC samples) represent object-attribute contexts. There is no object-attribute context for Bio parameters because each dataset is restricted to one value of one parameter (here IBGN red). `KBIOS` and `KPHCS` have no column since the samples are only described using the qualitative relations. The tables `RPHCS-ipb-PHCS`, `RBIOS-ipb-PHCS` , `RbPHC` and `RgPHC` represent object-object contexts. In these object-object contexts, a row is an object from the domain of the relation, a column is an object from the range of the relation and a cross indicates a link between two objects. For example, `RPHCS-ipb-PHCS` defines the temporal relations (`ipb`) between PhC samples and has `KPHCS` both as domain and range. `RbPHC` defines the qualitative relations between PhC samples and PhC parameters that have the blue (`b`) qualitative value.

Figure 2 represents the family of concept lattices obtained by applying RCA on the RCF illustrated in Tab. 2. There are three lattices, one for each formal context: $\mathcal{L}_{\texttt{KPHCS}}$ (PhC samples, Fig. 2(a)), $\mathcal{L}_{\texttt{KPHC}}$ (PhC parameters, Fig. 2(b)) and $\mathcal{L}_{\texttt{KBIOS}}$ (Bio samples, Fig. 2(c)). Each concept is represented by a box structured from top to bottom as follows: concept name, simplified intent and simplified extent. As said before, we have used the existential quantifier to build relational attributes. For instance, the intent of `C_KPHCS_2` from concept $\mathcal{L}_{\texttt{KPHCS}}$ contains the relational attribute ∃`RgPHC(C_KPHC_1)` inherited from concept `C_KPHCS_5`. This relational attribute is common to all PhC samples that measure a green PHOS parameter, which represents the extent of concept `C_KPHC_1` shown in Fig. 2(b).

Table 2: RCF composed of object-attribute contexts: KPHC, KBIOS and KPHCS; temporal object-object contexts: RBIOS-ipb-PHCS and RPHCS-ipb-PHCS; qualitative object-object contexts: RbPHC and RgPHC.

object-attribute contexts

KPHC	AZOT	PHOS
AZOT	×	
PHOS		×

KBIOS
S1_20/01
S1_28/12
S2_30/02

KPHCS
S1_17/01
S1_10/01
S1_25/12
S2_28/02
S2_20/02

object-object contexts

RBIOS-ipb-PHCS	S1_17/01	S1_10/01	S1_25/12	S2_28/02	S2_20/02
S1_20/01	×	×			
S1_28/12			×		
S2_30/02				×	×

RPHCS-ipb-PHCS	S1_17/01	S1_10/01	S1_25/12	S2_28/02	S2_20/02
S1_17/01		×			
S1_10/01					
S1_25/12					
S2_28/02					×
S2_20/02					

RbPHC	AZOT	PHOS
S1_17/01		
S1_10/01	×	
S1_25/12	×	
S2_28/02		
S2_20/02	×	

RgPHC	AZOT	PHOS
S1_17/01	×	×
S1_10/01		
S1_25/12		×
S2_28/02	×	×
S2_20/02		×

The navigation amongst the lattices shown in Fig. 2 follows the concepts used to build relational attributes. For example, the aforementioned relational attribute $\exists$RgPHC(C_KPHC_1) allows us to navigate from concept C_KPHCS_2 out of $\mathcal{L}_{\texttt{KPHCS}}$ to concept C_KPHC_1 out of $\mathcal{L}_{\texttt{KPHC}}$.

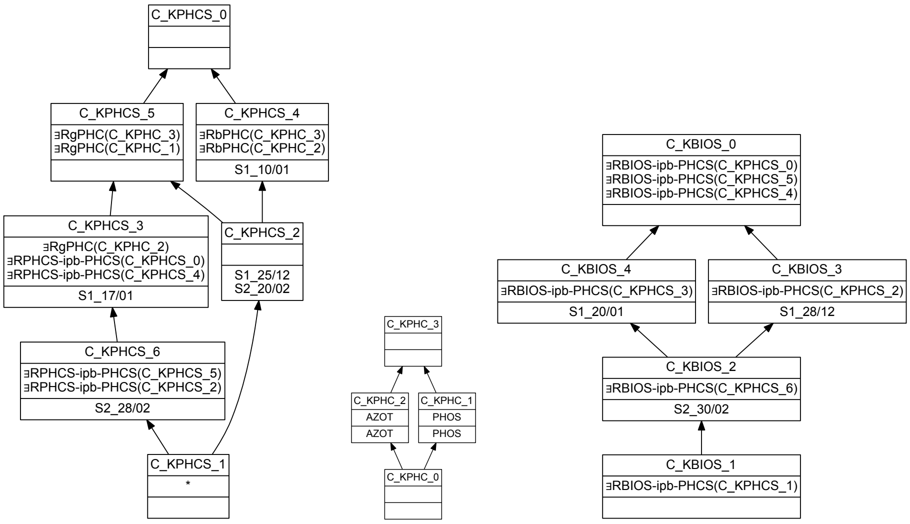
5 Measures of Interest for Temporal Concepts

To analyse the results of the RCA process, experts start from a *main lattice*, here the lattice $\mathcal{L}_{\texttt{KBIOS}}$, and navigate through the relational attributes linking concepts of different lattices. Besides, since RCA process can produce a large number of interrelated concepts, depending on the dataset volume and characteristics, some interestingness measures are required to select relevant concepts from where to start the navigation.

Such measures should take into account the specificity of concepts built on temporal objects, whereas well-known measures (e.g. concept stability) fit basic concepts. For example, Fig. 3 depicts two concept extents where the temporal objects are the Bio samples. Both concepts – that we call temporal concepts – have the same number of Bio samples and they cover the same geographical area. If two Bio samples are deleted, following the idea of stability measure, one of the site $S2$ and one of $S3$, then both concepts still have the same number of Bio samples but they cover different river sites.

To overcome this limitation, we introduce below an approach based on the distribution of temporal concept extents. The main idea in our method states that a concept is relevant if it is frequent and related to many sites where Bio samples are evenly distributed amongst these sites. Accordingly, we try to find temporal concepts whose intents represent universally available regularities in the studied geographical area. In our example, both concepts have the same frequency (7 samples), but the distribution is different: Concept_1 is more relevant than Concept_2.

Let (X, Y) be a formal concept of the main lattice, then its extent X is a set of temporal objects – or pairs – $(Object, Date)$. If the value of $Object$ is not identical for all the pairs, then the pairs can be grouped into categories by objects. We accordingly define $\bar{X}$ which represents the set of distinct objects



(a) $\mathcal{L}_{\texttt{KPHCS}}$ (b) $\mathcal{L}_{\texttt{KPHC}}$ (c) $\mathcal{L}_{\texttt{KBIOS}}$

Fig. 2: The family of concept lattices obtained by applying RCA on the RCF given in Tab.2. The $*$ symbol represents all the relational attributes of `KPHCS`.

from X pairs: $\bar{X} = \{o \in O | \exists t \in T, (o,t) \in X\}$, where O is the object set and T the set of dates.

Definition 1 (Absolute Frequency (ϕ_o)). *Let $C = (X,Y)$ be a temporal concept and o an object of $\bar{X}$. The absolute frequency of o in C, denoted ϕ_o, is equal to the number of distinct pairs of X where o occurs. $\bar{X}_\phi = \{(o, \phi_o)\, | o \in \bar{X}\}$.*

In our example (Fig. 3), $\bar{X_1} = \bar{X_2} = \{S1, S2, S3\}$. `Concept_1` has $\bar{X}_{1\phi} = \{(S1, 3)\, , (S2, 3)\, , (S3, 1)\}$ and `Concept_2` has $\bar{X}_{2\phi} = \{(S1, 5)\, , (S2, 1)\, , (S3, 1)\}$.

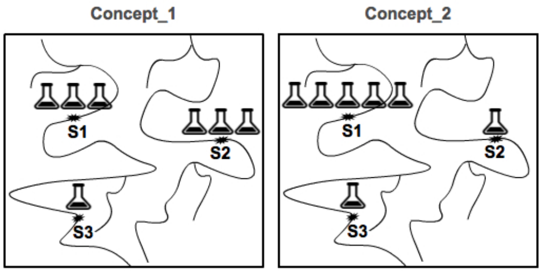


Fig. 3: Bio samples distribution by sites for two concept extents.

Definition 2 (Support and Richness (ρ)). *The support of a concept* (X, Y) *corresponds to the number of pairs* $(Object, Date)$ *out of* X*. Its richness, represented by* ρ*, is defined as the cardinality of* $\bar{X}$*.*

Definition 3 (Distribution index (IQV)). *The distribution of a concept* (X, Y) *describes the number of times each object out of* $\bar{X}$ *occurs in* X *and it is measured by the Index of Qualitative Variation (IQV, [10]). IQV is based on the ratio of observed differences in* $\bar{X}_\phi$ *to the total number of possible differences within* $\bar{X}_\phi$ $(\rho > 1)$*.*

$$IQV = \frac{\rho \left(|X|^2 - \sum_{i=1}^{\rho} {\phi_{o_i}}^2 \right)}{|X|^2 \, (\rho - 1)} \tag{1}$$

If $\rho = 1$*,* $IQV = 0$*.*

Our choice of IQV stems from the observation that the objects of $\bar{X}$ do not have an intrinsic ordering. Thus, measuring their distribution using the IQV [10] seems interesting. The IQV ranges from 0 to 1. When all pairs of X contain the same object, there is no diversity and the IQV is 0. In contrast, when there are different objects and all pairs of $\bar{X}_\phi$ have equal ϕ_o, there is even distribution and the IQV is 1.

Returning to our example (Fig. 3), both concepts have support $|X_1| = |X_2| = 7$ and richness $\rho_1 = \rho_2 = 3$. For `Concept_1` the distribution is $IQV_1 = \frac{3\left[7^2 - \left(3^2+3^2+1^2\right)\right]}{7^2(3-1)} = 0.91$ and for `Concept_2` $IQV_2 = 0.67$. Hence, `Concept_1` is computed as more relevant than `Concept_2` since its objects (Bio samples) are better distributed amongst the sites.

6 CPO-patterns for Helping Expert Analysis

Since our aim is to facilitate the analysis work, we propose, in addition to the selection of relevant concepts, to convert those concepts into cpo-patterns. Indeed cpo-patterns are structures with a graphical representation easy to read and understand (e.g. Fig. 4). The expert can choose a cpo-pattern that highlights interesting, surprising knowledge, and deepen the analysis by exploring the area in the lattice surrounding the corresponding concept. Thus, starting from the family of lattices built using RCA, we extract cpo-patterns following the approach proposed in [14]. It is worth pointing out that there is a cpo-pattern for each concept out of the lattice corresponding to the objects of interest for the study, i.e. $\mathcal{L}_{\texttt{KBIOS}}$ in our work.

Formally, let $\mathcal{I} = \{I_1, I_2, ..., I_m\}$ be a set of *items*. An *itemset* IS is a non empty, unordered, set of items, $IS = (I_{j_1} ... I_{j_k})$ where $I_{j_i} \in \mathcal{I}$. 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

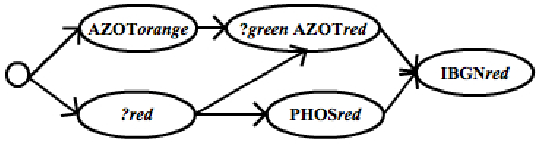


Fig. 4: Hybrid cpo-pattern: each vertex corresponds to a set of parameter values, edges represent the temporal relation, e.g. $\texttt{IBGN}_{red}$ is preceded by $\texttt{PHOS}_{red}$ that is preceded by $\texttt{?}_{red}$; this notation means that a PhC parameter with a red quality has been measured.

subsequences found in a sequence database. A *po-pattern* is a directed acyclic graph $G = (\mathcal{V}, \mathcal{E}, l)$. $\mathcal{V}$ is the set of vertices, $\mathcal{E}$ is a set of directed edges such that $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, and l is a labelling function mapping each vertex to an itemset. A partial order can be defined on G as follows: for all $\{u, v\} \in \mathcal{V}^2$, $u < v$ if there is a directed path from u to v. However, if there is no directed path from u to v, these elements are not comparable. Each path of the graph is a sequential pattern as defined before. The set of paths in G is denoted by $\mathcal{P}_G$. A po-pattern is associated to the set of sequences $\mathcal{S}_G$ that contains all paths of $\mathcal{P}_G$. Furthermore, let G and G' be two po-patterns with $\mathcal{P}_G$ and $\mathcal{P}_{G'}$ their sets of paths. G is a sub po-pattern of G', denoted by $G \preceq_g G'$, if $\forall M \in \mathcal{P}_G, \exists M' \in \mathcal{P}'_G$ such that $M \preceq_s M'$. A po-pattern G is closed, denoted *cpo-pattern*, if there exists no po-pattern G' such that $G \prec_g G'$ with $\mathcal{S}_G = \mathcal{S}_{G'}$.

As described in [14], thanks to the hierarchical structure of the RCA results, more or less accurate cpo-patterns are extracted. Based on their accuracy, three types of cpo-patterns could be defined: abstract, hybrid and concrete. Firstly, the *abstract* cpo-pattern represents an imprecise common trend of the analysed data. Secondly, the *hybrid* one, depicted in Fig. 4, corresponds to a more or less accurate common trend of the analysed data. Finally, the *concrete* cpo-pattern designates an accurate common trend of the analysed data.

7 Experiments and Discussion

The experiments are carried out on a MacBook Pro with a 2.9 GHz Intel Core i7, 8GB DDR3 RAM running OS X 10.9.5. RCA is applied using the RCAExplore[5] tool. For the extraction and selection of cpo-patterns we have developed an algorithm in Java 8 based on Java Collections Framework and Lambda Expressions.

Three sequential datasets (each dataset concerns only one Bio parameter having the *yellow* quality) from the Fresqueau project are analysed: IBD_{yellow}, IPR_{yellow} and $IBGN_{yellow}$. These datasets are interesting since the *yellow* quality of watercourses represents a median area between good ecological status and bad ecological status of watercourses. Other quality values have also been analysed but are not presented here. The objective is to extract more or less accurate

[5] http://dolques.free.fr/rcaexplore

Table 3: The results of mining the Fresqueau datasets. `Bio` and `PhC Samples` are the number of analysed samples; `Output` is the number of concepts from the main lattice ($\mathcal{L}_{\texttt{KBIOS}}$) and the lattice of PhC samples ($\mathcal{L}_{\texttt{KPHCS}}$); `CPO-patterns` is the number of the extracted cpo-patterns; `Execution Time` in seconds.

Datasets				RCA		Extraction			Execution Time
Index	Quality	Samples		Output		CPO-patterns			RCA & Extraction
		Bio	PhC	$\mathcal{L}_{\texttt{KBIOS}}$	$\mathcal{L}_{\texttt{KPHCS}}$	Concrete	Abstract	Hybrid	
IPR	yellow	80	194	35699	39605	433	3388	31877	593
IBD				32146	20947	503	1444	30198	115
IBGN				9414	11580	305	815	8293	32

cpo-patterns representing frequent PhC trends of watercourses common in many sites. To this end, the datasets are preprocessed and temporally modelled as described in Sections 3 and 4. The temporal relational analysis relies on the IceBerg algorithm [17], which result is a concept lattice of frequent closed itemsets. A 10% threshold is used only for the input of Bio samples (it corresponds to the lattice of Bio samples that covers the objects of interest from our work). The choice of this value allows us to focus on the cpo-patterns that describe many sites.

Table 3 shows some quantitative statistics regarding the temporal relational analysis and the extraction of cpo-patterns. The results in `Output` column show that the number of extracted concepts for the IBGN dataset is about 3 times smaller than the number of extracted concepts for the IPR and IBD datasets. This reveals greater heterogeneity in IPR and IBD datasets in contrast with IBGN. Consequently, cpo-patterns linking PhC and IBGN Bio parameters represent more examples and will provide more reliable forecasts of the *yellow* quality of watercourses.

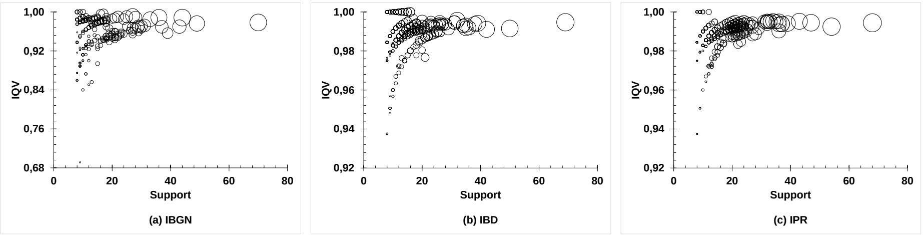
The `CPO-patterns` columns represent the different types of extracted cpo-patterns and illustrate their quite large number that has to be reduced. To this end, we select relevant cpo-patterns based on the support, richness and distribution of the associated concepts (see Section 5). Figure 5 shows three scatter-plots (for the three sets of extracted concrete cpo-patterns in Tab. 3) of the *distribution index (IQV)* with respect to the *support.* The diameter of the circles is proportional to the *richness.* The user can first explore a few selected cpo-patterns based on high thresholds for these measures. Then he/she can follow the cpo-pattern hierarchy to deepen the analysis, as described below, or select more cpo-patterns based on lower thresholds. For example, by defining two thresholds $\theta_{IQV} = 0.98$ and $\theta_{Support} = 25$, the top-6 (IBGN), the top-26 (IBD) and the top-30 (IPR) best distributed and most frequent cpo-patterns are selected. Focusing on IBD, if the thresholds are e.g. $\theta_{IQV} = 0.98$ and $\theta_{Support} = 20$, 52 cpo-patterns are selected. These cpo-patterns cover various numbers of sampling sites, and thus more or less extensive geographical areas. To select greater or smaller areas, the cpo-patterns are ranked by analysing the diameter of the circles.

The qualitative interpretation of the extracted cpo-patterns was performed by an hydroecologist. In Fig. 6 is an interesting excerpt from the main lattice of $IBGN_{yellow}$ dataset. This group of cpo-patterns is subsumed by the abstract cpo-pattern of C_KBIOS_868 (support = 28) that represents the less accurate common trend: *often before yellow IBGN are sampled simultaneously a green PhC parameter and another yellow PhC parameter*. Figure 6 also emphasizes the well-known correspondence between MOOX (organic matter pollutions) quality classes and IBGN ones: a *yellow* MOOX appears in the *yellow* IBGN cpo-pattern, which is associated to C_KBIOS_595. The concepts C_KBIOS_720, C_KBIOS_550 and C_KBIOS_400 highlight the impact of phosphorus pollution (PHOS) on macro-invertebrates (IBGN) that is a lesser-known fact.

Moreover, in Fig. 6 two benefits of exploring sequential data by means of RCA are observed. The first one is the generalisation order regarding the structure of the extracted cpo-patterns. For example, the structure of C_KBIOS_400 cpo-pattern is more specific than the structure of its ancestor cpo-patterns, i.e. there exist a projection from its ancestor cpo-patterns into C_KBIOS_400 cpo-pattern. The second benefit is the generalisation of items. For instance, the C_KBIOS_550 cpo-pattern reveals the rule $\{\texttt{PAES}_{green}, \texttt{PHOS}_{yellow}\} \rightarrow \{\texttt{IBGN}_{yellow}\}$ that is a specialisation of the rule revealed by the C_KBIOS_720 cpo-pattern, that is $\{?_{green}$, $\texttt{PHOS}_{yellow}\} \rightarrow \{\texttt{IBGN}_{yellow}\}$. These properties are useful for the expert who can navigate from specific to general patterns or vice versa.

8 Conclusion

We have introduced an original approach for exploring temporal data using RCA. Given a hydroecological dataset, where data represent Bio or PhC samples measured at a given time in a certain site, we find hierarchies of more or less general cpo-patterns that summarize the impact of PhC parameters on Bio ones. A comprehensive process for mining sequential datasets has been proposed: 1) preprocessing of the raw data based on domain knowledge, 2) relational analysis of the preprocessed data based on an original temporal data model, 3) selection of temporal concepts using the distribution, the richness and the support measures, and 4) extraction of cpo-patterns by navigating amongst temporal concepts (step



(a) IBGN

(b) IBD

(c) IPR

Fig. 5: Concrete cpo-patterns by distribution index, support and richness of the associated concepts.

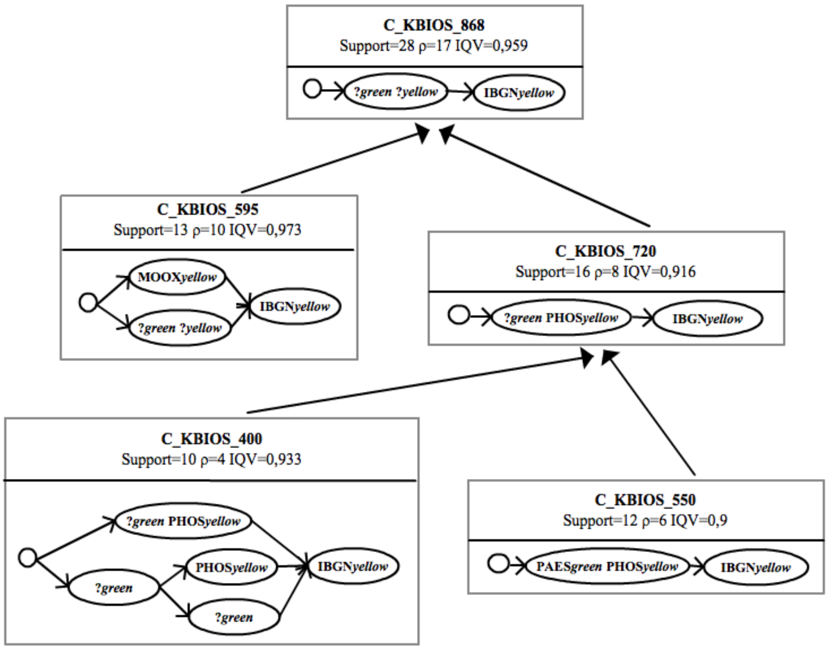


Fig. 6: Excerpt from a hierarchy of cpo-patterns (IBGN *yellow*).

detailed in [14]). Our method has been applied to sequential datasets from the Fresqueau project.

The main benefits of our approach are as follows. Using RCA produces hierarchical concepts, while cpo-patterns synthetize complex navigation paths, both facilitating the expert analysis. Furthermore, the proposed measures on temporal concepts are useful to select relevant information in our application.

In the future, we plan to apply our approach on other relational datasets. This will require to deeply investigate the behaviour of our measures and maybe to find other methods for selecting the extracted cpo-patterns.

References

1. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Int. Conference on Data Engineering. pp. 3–14 (1995)
2. Arévalo, G., Falleri, J.R., Huchard, M., Nebut, C.: Building abstractions in class models: Formal concept analysis in a model-driven approach. In: MoDELS 2006. pp. 513–527 (2006)
3. Bimonte, S., Boulil, K., Braud, A., Bringay, S., Cernesson, F., Dolques, X., Fabrègue, M., Grac, C., Lalande, N., Le Ber, F., Teisseire, M.: Un système décisionnel pour l'analyse de la qualité des eaux de rivières. Ingénierie des Systèmes d'Information 20(3), 143–167 (2015)

4. Buzmakov, A., Egho, E., Jay, N., Kuznetsov, S.O., Napoli, A., Raïssi, C.: On mining complex sequential data by means of FCA and pattern structures. International Journal of General Systems 45, 135–159 (2016)
5. Casas-Garriga, G.: Summarizing sequential data with closed partial orders. In: 2005 SIAM Int. Conference on Data Mining. pp. 380–391 (2005)
6. Džeroski, S.: Relational data mining. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 869–898. Springer (2005)
7. Fabrègue, M., Braud, A., Bringay, S., Grac, C., Le Ber, F., Levet, D., Teisseire, M.: Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. Ecological Informatics 24, 210–221 (2014)
8. Ferré, S.: The efficient computation of complete and concise substring scales with suffix trees. In: Formal Concept Analysis, pp. 98–113. Springer (2007)
9. Ferreira, C.A., Gama, J., Costa, V.S.: Exploring multi-relational temporal databases with a propositional sequence miner. Progress in AI 4(1-2), 11–20 (2015)
10. Frankfort-Nachmias, C., Leon-Guerrero, A.: Social Statistics for a Diverse Society, chap. Measures of Variability. SAGE Publications (2010)
11. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer (1999)
12. Klimushkin, M., Obiedkov, S., Roth, C.: Approaches to the selection of relevant concepts in the case of noisy data. In: Formal Concept Analysis, pp. 255–266. Springer (2010)
13. Kuznetsov, S.O.: On stability of a formal concept. Annals of Mathematics and Artificial Intelligence 49(1-4), 101–115 (2007)
14. Nica, C., Braud, A., Dolques, X., Huchard, M., Le Ber, F.: Extracting Hierarchies of Closed Partially-Ordered Patterns using Relational Concept Analysis. In: International Conference on Conceptual Structures, ICCS'2016, Annecy, France. pp. 1–14. Springer (2016)
15. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G.: A Method based on Temporal Concept Analysis for Detecting and Profiling Human Trafficking Suspects. In: Artificial Intelligence and Applications, AIA 2010, Innsbruck, Austria. pp. 1–9 (2010)
16. 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)
17. Stumme, G.: Efficient data mining based on formal concept analysis. In: Database and Expert Systems Applications, pp. 534–546. Springer (2002)
18. Wolff, K.E.: Temporal Concept Analysis. In: ICCS-01 Workshop on Concept Lattice for KDD, 9th Int. Conference on Conceptual Structures. pp. 91–107 (2001)