Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment

Mickaël Fabrègue, Agnès Braud, Sandra Bringay, Corinne Grac, Florence Le Ber, Danielle Levet, Maguelonne Teisseire

Ecological Informatics 24 (2014) 210–221

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment

Mickaël Fabrègue [a,b,*], Agnès Braud [c], Sandra Bringay [d], Corinne Grac [e], Florence Le Ber [b], Danielle Levet [f], Maguelonne Teisseire [a]

[a] TETIS, IRSTEA, Montpellier, France
[b] ICube, University of Strasbourg/ENGEES, CNRS, Illkirch, France
[c] ICube, University of Strasbourg, CNRS, Illkirch, France
[d] LIRMM, Montpellier 3 University, CNRS, France
[e] LIVE, University of Strasbourg/ENGEES, CNRS, France
[f] AQUASCOP, Technopole d'Angers, Beaucouze, France

ARTICLE INFO

Article history:
Received 17 June 2014
Received in revised form 30 August 2014
Accepted 2 September 2014
Available online 16 September 2014

Keywords:
Data mining
Temporal patterns
Discriminant patterns
Hydro-ecology
River quality

ABSTRACT

We propose a new data mining process to extract original knowledge from hydro-ecological data, in order to help the identification of pollution sources. This approach is based (1) on a domain knowledge discretization (quality classes) of physico-chemical and biological parameters, and (2) on an extraction of temporal patterns used as discriminant features to link physico-chemistry with biology in river sampling sites. For each bio-index quality value, we obtained a set of significant discriminant features. We used them to identify the physico-chemical characteristics that impact on different biological dimensions according to their presence in extracted knowledge. The experiments meet with the domain knowledge and also highlight significant mismatches between physico-chemical and biological quality classes. Then, we discuss about the interest of using discriminant temporal patterns for the exploration and the analysis of temporal environmental data such as hydro-ecological databases.

1. Introduction

Identifying pollution sources in aquatic ecosystems is currently a major research area and remains a complex task. Many parameters are involved in the determination of aquatic ecosystems quality. These parameters are related to different aspects, such as biology, physico-chemistry and hydromorphology. The importance of having operational tools to help in the interpretation of complex information concerning the water quality of rivers and their functioning, as well as assessment of the effectiveness of ongoing action programs is underlined by international directives such as the European Water Framework Directive (E. Union, 2000). Therefore, it is important to propose new methods that take into account the complexity of the problem.

Measures of these different aspects are performed in river stations by several organizations, with specific research goals. Because the data collected by each actor of the domain have become substantial, it is important to design and implement a large, common and consistent database to aggregate these complementary data. In order to meet this issue, the French ANR[1] Fresqueau project[2] has begun in 2011. This project aims at collecting and unifying databases that are linked to the quality of water bodies. They include biological, physico-chemical and also hydro-morphological data. The result is a consistent spatio-temporal database that brings together information related to north-east and south-east French watersheds. It concerns 11,329 sampling sites spread over 161,100 km^2 that represent 29.45% of metropolitan France. These watersheds are grouped into two major hydrographic areas which are *Rhin-Meuse* (north-east), denoted as *RM* and *Rhône Méditerranée Corse* (south-east), denoted as *RMC*. Fig. 1 illustrates their respective geographic scopes. The dark gray area corresponds to *RM* while the black area corresponds to *RMC*. White line separations in the figure correspond to the different watershed delimitations.

Several dimensions of analysis have been collected. Fig. 2 illustrates the dimensions of analysis of the database: physico-chemistry, hydrobiology, climate, land use, hydrology and hydromorphology. The 11,329 sampling sites are described by these 6 dimensions. The objective is to provide researchers with a maximum amount of data to analyze. It aims at facilitating studies that focus on the relations between various environmental aspects, or the impact of one aspect on another one. Furthermore, some of these different environmental aspects involve a temporal dimension. For example, physico-chemical parameters may be sampled every two months in sampling sites. Considering the different hydro-ecological parameters with their temporal dimension allows the application of original methods. Indeed, some temporal

* Corresponding author.
E-mail address: mickael.fabregue@teledetection.fr (M. Fabrègue).
[1] French National Research Agency.
[2] http://engees-fresqueau.unistra.fr/.

http://dx.doi.org/10.1016/j.ecoinf.2014.09.003
1574-9541/

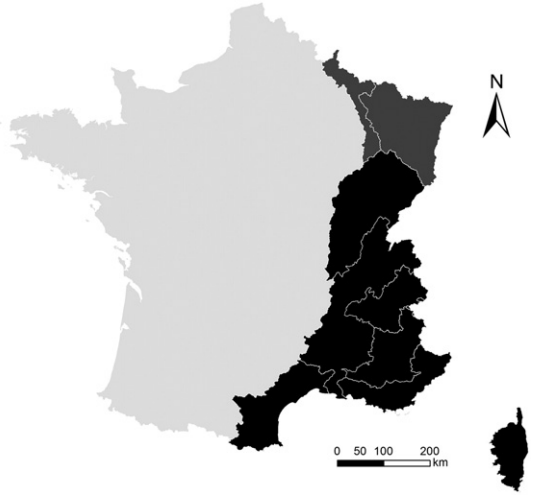


Fig. 1. French watersheds concerned by the Fresqueau database.

data mining approaches are well-adapted to tackle such issues. With specific data structures, they are able to process temporal information that describe environmental aspects.

This paper presents such a data mining method applied on hydrobiological and physico-chemical aspects. It addresses the following issue:

Can we temporally link sets of physico-chemical parameter values with bio-index values?

Identifying these links is important to evaluate more precisely the impact of physico-chemistry on biology. Finding temporally ordered sets of physico-chemical parameter values may help to highlight the synergy produced by their combination. The presented method proposes an original temporal pattern based approach, called discriminant closed partially ordered patterns, to obtain these correlations.

In Section 2, we present existing approaches from the literature. In Section 3, we describe our method divided into three parts:

1. Section 3.1 details the preprocess operations performed on the dataset, and the construction of quality class sub-datasets that correspond to each bio-index.
2. Mining discriminant partially ordered patterns is presented in Section 3.2.
3. The last part consists in selecting and reducing the discriminant partially ordered pattern result set (Section 3.3).

We then provide experimental results performed on the Fresqueau dataset (Section 4) and we finish with a discussion section (Section 5). Fig. 3 synthesizes this process, which is detailed in Section 3.

2. Related work

Several works investigated the task of mining hydrological data.

An important amount of studies focus on macro-invertebrate communities (D'heygere et al., 2003; Dakou et al., 2007; Dedecker et al., 2004; Goethals et al., 2007). For example, Dakou et al. (2007) used decision tree models in order to predict the habitat suitability of some macro-invertebrate taxa in river Axios (Greece). Authors considered physico-chemical and structural characteristics of the river. With the same goal, the efficiency of artificial neural networks in predicting macro-invertebrate taxa in Zwalm (Belgium) river has been shown by Dedecker et al. (2004).

The impact of hydrologic alterations on fish communities in Illinois River has been identified by Yang et al. (2008). Based on 32 indicators of hydrologic alteration, authors highlight the most ecologically relevant indicators by using a genetic programming approach.

Some other authors focused on flora instead of fauna. The first comprehensive checklist of diatoms (948 taxa) with ecological indicator values for pH, salinity, nitrogen uptake metabolism, saprobity, trophic state and moisture was presented by Van Dam et al. (1994). Recently, the physico-chemical impact on diatom communities has been studied by Kocev et al. (2010). They used a multi-target regression trees approach and identified a significant impact of metallic ions and nutrients on diatoms. Recknagel et al. (2013) analyzed phytoplankton phyla populations in Lake Kinneret (Israel), by using a hybrid evolutionary algorithm. Authors showed that considering both physico-chemical and biological variables in models provides the best results in the prediction of population dynamics. Likewise, Bertaux et al. (2009) rely on Formal Concept Analysis to study biological traits of macrophytes taxa in Rhin-Meuse watershed. The goal is to link environmental variables with biological trait granularity in order to identify groups of taxa adapted to a particular environmental context.

State-of-the-art methods show the importance of considering and combining biological and physico-chemical variables in order to find relevant knowledge. Nevertheless, none of these studies has taken into account the temporal aspect based on temporal pattern mining approaches, which is relevant to analyze pollution dynamics. The approach presented in this study is well-adapted to temporal datasets

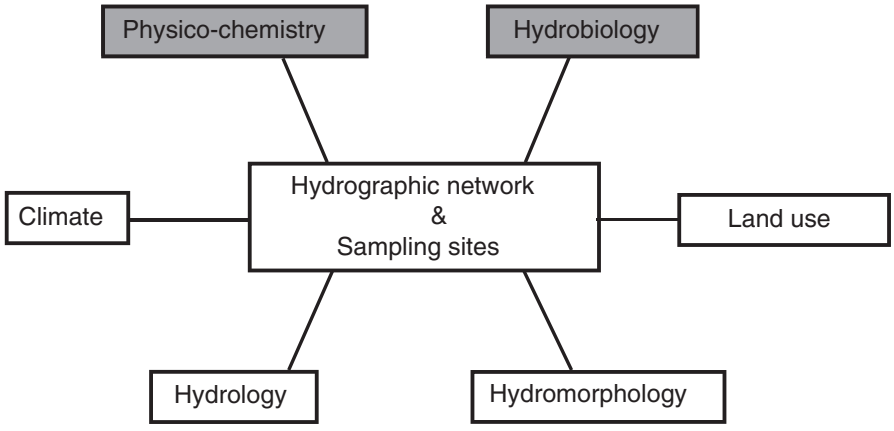


Fig. 2. Categories of data in the Fresqueau database.

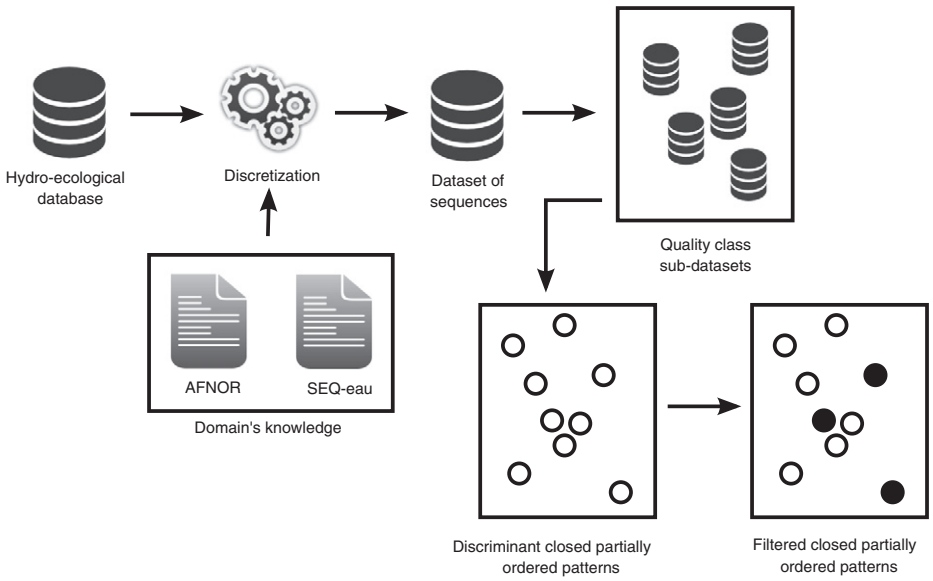


Fig. 3. Process illustration.

with multiple variables represented here by biology and physico-chemistry. Furthermore, extracted knowledge from temporal pattern approaches is easy to analyze by experts. We now introduce temporal pattern mining approaches.

2.1. Temporal pattern mining approaches

Currently, the most common pattern-based tools used to explore temporal data are sequential pattern mining approaches. In the literature, such pattern approaches have been widely used in many studies like analysis (Geng and Hamilton, 2006), classification (Cheng et al., 2007, 2008) or prediction (Wang et al., 2008). They have been first introduced by Agrawal and Srikant (1995) and are a temporal extension of association rules (Agrawal and Srikant, 1994) that have been first developed and used to find strong correlations among supermarket products. Sequential patterns are more complex than association rules since they lead to a more important search space. They are used when information is totally ordered according to a specific criterion, which is most often temporal. Let us consider a temporal dataset and a sequential pattern $\langle(\textit{Low oxygen level})(\textit{Disappearance of species})\rangle$: 30 % extracted from the transactions of this dataset. This sequential pattern means that the *Low oxygen level* event is temporally followed by the *Disappearance of species* event with a frequency of 30% in the dataset. A transaction is represented by a sequence of elements ordered on the temporal dimension. Mining such features according to the temporal aspect is very useful for specialists in various domains such as software engineering (Ren et al., 2009), medicine (Sallaberry et al., 2011) and marketing (George and Binu, 2012). Despite their advantages, sequential patterns often bring limited information since they only provide totally ordered information about data. To illustrate this, let us consider a second pattern $\langle(\textit{Presence of pesticides})(\textit{Disappearance of species})\rangle$: 30 % discovered in the same dataset. It is possible to extract the two patterns exactly from the same set of transactions: they coexist in the dataset. The coexistence of sequential patterns is not taken into account with this method. However, this coexistence can be synthesized based on partial ordering. Fig. 4 presents a so-called partially ordered pattern that combines the two previous sequential patterns.

This partially ordered pattern means that the *Disappearance of species* event is frequently preceded by two events *Low oxygen level* and *Presence of pesticides*, which themselves are not ordered. Partially ordered pattern approaches used on hydrobiological data have some advantages:

1. They are well-adapted to the temporal aspect of the dataset.
2. They provide more information on order among elements than sequential patterns.
3. They are represented as a directed acyclic graph that facilitates the understanding, which is important for hydrobiologists.

3. Material and methods

This work focuses on partially ordered patterns and more precisely on discriminant closed partially ordered patterns, denoted by DCPO-patterns in the following. The discriminative property of patterns had been studied in itemset mining (Cheng et al., 2007, 2008). It is related to the interestingness measure domain that consists in applying statistical measures on patterns in order to select the most interesting ones according to analyst's needs. Geng and Hamilton (2006) work is an exhaustive survey on existing interestingness measures. These measures can be applied on all kinds of patterns since they are mainly based on their frequency. Nevertheless, existing measures are mainly efficient in the case of binary classes. In the following, we propose an interestingness measure that considers the case of n classes with the aim of retrieving the most significant patterns for each class, i.e. the DCPO-patterns that are more frequent in the considered class than in others. As we shall see later, we construct such classes, denoted as quality class sub-datasets, based on bio-index values. To simplify, the aim is to extract DCPO-patterns that are more frequent in a polluted quality class than in a non-polluted one, or conversely.

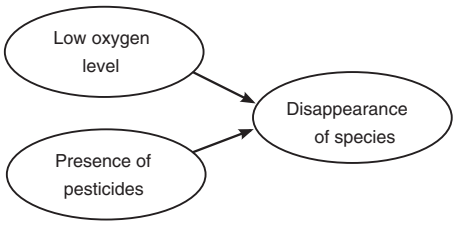


Fig. 4. Example of partially ordered pattern with a frequency of 30%.

3.1. Data and preprocessing

Before extracting DCPO-patterns, we have to apply different preprocessing steps to the data. The data focus on sampling stations, which are characterized by some information such as their spatial coordinates and a list of characteristic values sampled at various time intervals. We detail this in the following. To facilitate the understanding of the overall process, we take as example two sampling stations located along a same river and illustrated by Fig. 5.

Since our approach is temporal-based, we have to consider river sampling timestamps. The sampled characteristics are numerous and varied, they are divided into two major categories representing the physico-chemistry on the one hand, and the biology on the other hand.

Biological data

They concern the flora and the fauna taxa living in the river. They are divided in several biological dimensions that are macro-invertebrates, fishes, macrophytes and diatoms. Each dimension is represented by a bio-index, giving a global quality mark about the viability of the hydro-ecosytem for this dimension. Bio-indices are based on French normalized standards. Bio-indices used in this paper are IBGN (AFNOR (Association Française de NORmalisation), 1992, révision 2004) (macro-invertebrates), IBD (AFNOR (Association Française de NORmalisation), 2000, révision 2007) (diatoms) and IPR (AFNOR (Association Française de NORmalisation), 2004) (fishes) bio-indices.

Physico-chemical data

Physico-chemistry is the measurement of various physico-chemical parameters. We can mention for example the measurement of temperature, oxygen levels, analysis of elements such as nitrates and phosphorus or the presence of molecules such as synthetic pesticides and hydrocarbons. As we explain below in the preprocess of the data, these parameters are treated by family rather than individually because of their large number. The Fresqueau dataset gathers more than 900 physico-chemical parameters.

Table 1 gives an example of biological and physico-chemical samplings on the two sampling stations of Fig. 5. Samples of *Site* 1 correspond to the temporal period from February 2007 to July 2008, and samples in *Site* 2 correspond to the temporal period from January 2004 to August 2005. This example provides sampling of five physico-chemical parameters and measures of the IBGN bio-index at different timestamps. Briefly, ammonium (NH_4^+), Kjeldahl nitrogen (NKJ) and nitrite (NO_2^-) are a part of nitrogenous matter. Orthophosphate (PO_4^{3-}) and total phosphorus (P) are representative of the level of phosphorous matter in water. In our illustrative example, we only consider the IBGN bio-index which characterizes the macro-invertebrate dimension. Some macro-invertebrate taxa are for instance typical of a good river quality, denoted as polluo-sensitive, while some other taxa are not. Then, a high abundance of polluo-sensitive taxa often leads to a good IBGN score. These biological samplings are done once a year for each site. In the dataset, a value of 2.331 mg/l NH_4^+ is for example sampled on July 2004 for *Site* 2 and an IBGN score of 8/20 is measured two months later on September 2004 in the same sampling site.

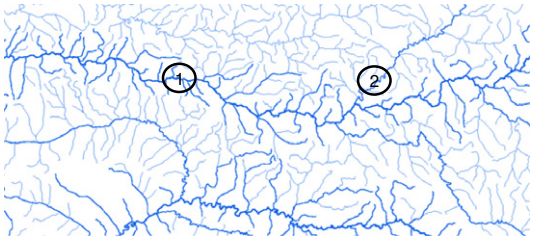


Fig. 5. Hydro-ecological network example.

Table 1
Dataset example.

Site	Date	NH_4^+	NKJ	NO_2^-	PO_4^{3-}	P	$IBGN_{/20}$
Site 1	02/07	–	–	–	0.123	0.032	–
	06/07	–	0.672	0.026	–	–	–
	07/07	0.088	1.235	0.134		0.011	–
	09/07	–	–	–	–	–	17
	12/07	0.154	–	0.246	0.168	0.338	–
	02/08	0.062	0.040	0.091	0.025	0.003	–
	04/08	–	0.023	0.198	–	–	–
	05/08	–	–	–	–	–	12
	07/08	–	–	–	0.046	0.009	–
Site 2	01/04	0.043	0.146	0.421	–	–	–
	04/04	–	–	–	1.325	0.093	–
	07/04	2.331	7.993	0.252	0.132	0.266	–
	08/04	–	1.414	–	–	–	–
	09/04	–	–	–	–	–	8
	11/04	0.117	0.0844	–	0.688	–	–
	12/04	–	–	–	0.067	0.278	–
	03/05	–	0.182	0.0310	0.137	–	–
	06/05	0.004	–	0.012	0.035	0.134	–
	08/05	–	–	–	–	–	10

The Fresqueau dataset only contains numerical values. Pattern-based methods only perform on discrete data. We now present the discretization process based on domain knowledge.

3.1.1. Data discretization

To process the dataset with pattern-based approaches, each physico-chemical or biological variable needs to be discretized. Instead of choosing arbitrary intervals, we base our discretization on existing works in hydrobiology. Indeed, French water agencies published technical reports that provide quality intervals for biology and physico-chemistry. For both parameter categories, there are five quality values related to the river quality: “Very good”, “Good”, “Medium”, “Bad” and “Very bad” represented by colors *Blue*, *Green*, *Yellow*, *Orange* and *Red*, respectively. In the following, a quality value for a parameter is called a quality class. The bio-index quality classes are given by AFNOR standards of bio-indices (AFNOR (Association Française de NORmalisation), 1992, révision 2004 for IBGN, AFNOR (Association Française de NORmalisation), 2000, révision 2007 for IBD and AFNOR (Association Française de NORmalisation), 2004 for IPR).

Although biological quality classes are easy to apply on our data, physico-chemical quality classes require a more complex procedure.

Biological discretization

AFNOR standards provide discretization quality intervals for many bio-indices and in particular the three bio-indices focused in this paper: IBGN, IPR and IBD. Table 2 gives us discretization intervals for the IBGN bio-index. It means that an IBGN value of 15 is discretized as a *Green* quality and an IBGN value of 6 is discretized as an *Orange* quality.

Physico-chemical discretization

The number of physico-chemical parameters is huge. The SEQ-eau Standard allows us to significantly reduce this number of parameters by providing 15 macro-parameters that group the initial ones. These macro-parameters are relevant since they group parameters in families according to their nature or their function (as phosphorus, organic matters, pesticides, …). Given a macro-parameter, the process consists in computing the discretized quality class for each included parameter to

Table 2
Discretization thresholds for the IBGN bio-index according to NFT90-350 standardized method (AFNOR (Association Française de NORmalisation), 1992, révision 2004).

Parameter	Blue	Green	Yellow	Orange	Red
IBGN	[20,17]	]17,13]	]13,9]	]9,5]	]5,0]

Table 3
Quality class intervals composing AZOT and PHOS macro-parameters according to SEQ-eau.

Group	Parameter	Blue	Green	Yellow	Orange	Red
AZOT	NH_4^+ (mg/l)	[0,0.1[	[0.1,0.5[	[0.5,2[	[2,5[	[5,∞]
	NKJ (mg/l)	[0,1[	[1,2[	[2,4[	[4,10[	[10,∞]
	NO_2^- (mg/l)	[0,0.03[	[0.03,0.3[	[0.3,0.5[	[0.5,1[	[1,∞]
PHOS	PO_4^{3-} (mg/l)	[0,0.1]	[0.1,0.5[	[0.5,1[	[1,2[	[2,∞]
	P (mg/l)	[0,0.05[	[0.05,0.2[	[0.2,0.5[	[0.5,1[	[1,∞]

assign the worst quality class to the macro-parameter. Table 3 provides the two macro-parameters related to the five physico-chemical parameters illustrated in the initial dataset in Table 1. For example with an orthophosphate (PO_4^{3-}) value of 0.026 and a total phosphorus value of 0.67, PO_4^{3-} is discretized as a *Blue* quality class and total phosphorus is discretized as a *Yellow* quality class. The macro-parameter PHOS is then discretized as a *Yellow* quality class by taking the worst quality in the included macro-parameters. Furthermore, it is important to note that a macro-parameter can be computed even if there are missing values for some included parameters, at least one valued parameter is required. The reason is that some macro-parameters contain a large number of parameters, and the sampling of some of these parameters is expensive and is rarely carried out in rivers. For example, let us take samples on April 2008 from *Site* 1 (Table 1), there are values for NKJ and NO_2^- parameters but the value is missing for NH_4^+ parameter. Given the discretization table, NKJ value is discretized as a *Blue* quality class and NO_2^- value as a *Green* quality class, then the macro-parameter AZOT value has a *Green* quality class.

Based on the initial dataset in Table 1, Table 4 gives us the discretized dataset obtained by grouping parameters. We can note that the number of variables decreases in this new dataset: based on Table 3, NH_4^+, NKJ and NO_2^- parameters are reduced to the AZOT macro-parameter while PO_4^{3-} and P parameters are reduced to the PHOS macro-parameter.

Applying AFNOR and SEQ-eau standards has several advantages:

1. The discretization process is based on domain knowledge instead of an arbitrary number of intervals.
2. It significantly reduces the initial dataset variables by providing macro-parameters.
3. It avoids some missing values since only one variable is needed to be able to compute the corresponding macro-parameter.

The dataset is now ready to be transformed in sequences.

Table 4
Discretized dataset example.

Site	Date	AZOT	PHOS	IBGN
Site 1	02/07	–	Green	–
	06/07	Blue	–	–
	07/07	Green	Blue	–
	09/07	–	–	Blue
	12/07	Green	Yellow	–
	02/08	Green	Blue	–
	04/08	Green	–	–
	05/08	–	–	Yellow
	07/08	–	Blue	–
Site 2	01/04	Yellow	–	–
	04/04	–	Orange	–
	07/04	Orange	Yellow	–
	08/04	Green	–	–
	09/04	–	–	Orange
	11/04	Green	Yellow	–
	12/04	–	Yellow	–
	03/05	Green	Green	–
	06/05	Blue	Green	–
	08/05	–	–	Yellow

3.1.2. Sequence preprocessing

The aim of this step is to build sequences by ordering measure samplings according to their timestamp. Pattern mining approaches consider items, which would be in our case the physico-chemical and biological measures. An itemset *IS* is a non-ordered group of measures sampled at a same timestamp. A sequence $S = \langle IS_1 IS_2 \ldots IS_{|S|} \rangle$ is a non-empty and ordered list of itemsets, i.e. groups of measures ordered according to their timestamp. For each river site, a sequence is built according to all discretized variables of the dataset in Table 4. Table 5 presents these sequences.

Let us consider the river *Site* 1, there are a *Green* quality class PHOS at timestamp 02/07, a *Blue* quality class AZOT at timestamp 06/07, a *Green* quality class AZOT and a *Blue* quality class PHOS at timestamp 07/07. Thus *Site* 1 sequence starts with the sequence $\langle (\text{PHOS}^{Green})(\text{AZOT}^{Blue})(\text{AZOT}^{Green}, \text{PHOS}^{Blue}) \rangle$. It means that item PHOS^{Green} is temporally followed by item AZOT^{Blue}, itself followed by the two items AZOT^{Green} and PHOS^{Blue}. Given the dataset of sequences, we now present the next step that cuts sequences in order to construct specific biological quality class sub-datasets.

3.1.3. Quality class sub-datasets

At this step, the obtained sequence dataset is not yet easily explorable to link physico-chemistry with biology. Site sequences are often long since they cover many years of samplings. For a same sampling site, a good biological value may be measured at a given timestamp and a bad biological value may be measured at another timestamp. Thus, a sampling site sequence may describe different water quality episodes over the time. Furthermore, given a biological measured value, considering physico-chemical values measured some years before is not relevant. Then a time-constraint is necessary to avoid or limit non-sense knowledge.

To tackle those issues the idea is as follows, the sequence dataset is transformed into quality class sub-datasets composed of sequences for each bio-index, where each sequence represents a water quality episode from sequences of sampling stations. A quality class sub-dataset corresponds to a given quality class of a bio-index, for example the value green for the IBGN. It means that for IBGN bio-index, there are five quality class sub-datasets representing a different discretized IBGN value (*Blue*, *Green*, *Yellow*, *Orange* or *Red*). In the dataset in Table 5, discretized bio-index values are included as variables in sequences. Then to manage bio-indices to build different datasets, we have to cut initial sequences in sub-sequences for each bio-index quality value. For example, given the value IBGN^{Blue}, it consists in collecting all sub-sequences in the dataset of sequences that precede an item IBGN^{Blue} given a chosen time interval represented by a window.

We illustrate this step by cutting sequences from the discretized dataset in Table 4. In this dataset, the IBGN bio-index has three different values: *Blue*, *Yellow* and *Orange*. Thus we obtain three quality class sub-datasets IBGN^{Blue}, IBGN^{Yellow} and IBGN^{Orange}. The selected temporal interval is equal to 6 months. It collects all discretized physico-chemical variables that appear in this interval before a discretized bio-index quality class. This process is illustrated in Table 6 where biological index values are framed, and not collected variables are blurred. For example, the first PHOS^{Green} item from *Site* 1 sequence is not collected since it is sampled 7 months before the first IBGN measure. Then according to this process, we build the sub-datasets in Table 7, where each IBGN quality class is composed of the set of sub-sequences that precede its measurement. For easier understanding of the following section with a more complete example, we add some new sub-sequences in Table 7 that do not exist in Table 5.

We can now use this IBGN dataset to mine DCPO-patterns. The following section presents our temporal pattern based method.

Table 5
Sequence dataset example.

Site	Sequence
Site 1	$\langle(\text{PHOS}^{Green})(\text{AZOT}^{Blue})(\text{AZOT}^{Green},\text{PHOS}^{Blue})(\text{IBGN}^{Blue})(\text{AZOT}^{Green},\text{PHOS}^{Yellow})(\text{AZOT}^{Green},\text{PHOS}^{Blue})(\text{AZOT}^{Green})(\text{IBGN}^{Yellow})(\text{PHOS}^{Blue})\rangle$
Site 2	$\langle(\text{AZOT}^{Yellow})(\text{PHOS}^{Orange})(\text{AZOT}^{Orange},\text{PHOS}^{Yellow})(\text{AZOT}^{Green})(\text{IBGN}^{Orange})(\text{AZOT}^{Green},\text{PHOS}^{Yellow})(\text{PHOS}^{Yellow})(\text{AZOT}^{Green},\text{PHOS}^{Green})(\text{AZOT}^{Blue},\text{PHOS}^{Green})(\text{IBGN}^{Yellow})\rangle$

Table 7
Quality class sub-datasets example for IBGN bio-index.

Sub-datasets	Sequence
IBGN^{Blue}	$\langle(\text{AZOT}^{Blue})(\text{AZOT}^{Green},\text{PHOS}^{Blue})\rangle$
	$\langle(\text{AZOT}^{Blue},\text{PHOS}^{Green})(\text{PHOS}^{Green})(\text{AZOT}^{Yellow},\text{PHOS}^{Blue})\rangle$
IBGN^{Yellow}	$\langle(\text{AZOT}^{Green},\text{PHOS}^{Green})(\text{AZOT}^{Blue},\text{PHOS}^{Green})\rangle$
	$\langle(\text{PHOS}^{Orange})(\text{AZOT}^{Orange},\text{PHOS}^{Yellow})(\text{AZOT}^{Green},\text{PHOS}^{Yellow})\rangle$
	$\langle(\text{AZOT}^{Green},\text{PHOS}^{Yellow})(\text{AZOT}^{Green},\text{PHOS}^{Blue})(\text{AZOT}^{Green})\rangle$
IBGN^{Orange}	$\langle(\text{AZOT}^{Orange},\text{PHOS}^{Yellow})(\text{AZOT}^{Red},\text{PHOS}^{Orange})(\text{AZOT}^{Green},\text{PHOS}^{Yellow})\rangle$
	$\langle(\text{PHOS}^{Orange})(\text{AZOT}^{Orange},\text{PHOS}^{Yellow})(\text{AZOT}^{Green})\rangle$

3.2. Pattern extraction

This step is the core of our proposition. It consists in extracting closed partially ordered patterns by quality class sub-dataset. In the general case, extracting closed partially ordered patterns is a complex task, because on real datasets the search space is often huge. It is related to combinatorial problems and pattern extraction approaches are all based on the same fundamentals. We use the algorithm *OrderSpan* that we proposed in Fabrègue et al. (2013) and we adapted it for mining multiple quality class sub-datasets. Initially, *OrderSpan* is only able to mine a single dataset. The trick is to mine iteratively DCPO-patterns for each quality class sub-datasets, while the non-discriminant DCPO-patterns are removed in a post process, as detailed below.

A pattern is characterized by a support, or a frequency, that represents the number of sequences in which the pattern appears. In the case of closed partially ordered patterns, a pattern appears in a sequence (or is supported by a sequence) if the order between all elements in the pattern is also observed in the sequence. Let us consider the partially ordered pattern in Fig. 6. This partially ordered pattern is supported by all sequences from quality class sub-dataset IBGN^{Orange} and none from quality class sub-dataset IBGN^{Blue}. Indeed, in all sequences from quality class sub-dataset IBGN^{Orange}, items AZOT^{Orange} and PHOS^{Orange} both appear and are followed by item AZOT^{Green}. Conversely, AZOT^{Orange} and PHOS^{Orange} are ordered differently in the sequences. Thus, this pattern has a support of 2 and a frequency of 2/2 in quality class sub-dataset IBGN^{Orange}.

In order to limit the search space exploration, a minimum frequency parameter noted θ is required as algorithm parameter. Given a θ value, it means that all extracted partially ordered patterns have a frequency higher than θ. In our case, we have to mine multiple quality class sub-datasets in parallel since each quality class sub-dataset is considered as an independent sequence dataset. It consists in extracting all partially ordered patterns whose support is higher than a minimum frequency θ in the concerned quality class sub-dataset. In addition, we retrieve the frequency of partially ordered patterns in other quality class sub-datasets. Then, each mined partially ordered pattern has multiple frequencies. For example, the partially ordered pattern in Fig. 7 has a frequency of 0/2 (support of 0) in quality class sub-dataset IBGN^{Blue}, 1/3 in quality class sub-dataset IBGN^{Yellow} and 2/2 in quality class sub-dataset IBGN^{Orange}. This pattern can be mined from IBGN^{Yellow} quality class sub-dataset with $\theta \leq 1/3$ or from IBGN^{Orange} quality class sub-dataset with $\theta \leq 2/2$.

Furthermore, we only consider partially ordered patterns that are closed, i.e. CPO-patterns. The closure property allows us to extract a smaller set of patterns without information loss. This property works on the following principle: given a partially ordered pattern P and $\mathcal{S}$ the set of sequences that support it, if there is no other partially ordered pattern P' such that P' is more specific than P and P' supported by all sequences in $\mathcal{S}$, then P is closed. More details about this property are provided in Fabrègue et al. (2013). For example, the partially ordered pattern in Fig. 6 is not closed since we point out that each time this pattern is supported by a sequence, at the same time we observe a PHOS^{Yellow} item. It leads to the partially ordered pattern in Fig. 7 that includes the pattern in Fig. 6 which has exactly the same support or frequency in each quality class sub-dataset. Thus, the partially ordered pattern in Fig. 6 is not closed and is redundant with respect to the partially ordered pattern in Fig. 7.

3.2.1. Discriminant patterns filtering

This step is a filtering process on extracted partially ordered patterns. For each quality class sub-dataset, not all extracted closed partially ordered patterns are meaningful and relevant. We opted to adapt the *Growth Rate* interestingness measure. *Growth Rate* had been first used in emerging pattern mining (Dong and Li, 1999) and is defined by Definition 1.

Definition 1. *Given a closed partially ordered pattern P and two quality class sub-datasets $\mathcal{C}_1$ and $\mathcal{C}_2$, the growth rate of P in $\mathcal{C}_1$ with respect to $\mathcal{C}_2$, denoted $GR(P, \mathcal{C}_1, \mathcal{C}_2)$, is defined as*

$$\begin{cases} 0, & \text{if } Freq_{\mathcal{C}_1}(P) = 0 \text{ and } Freq_{\mathcal{C}_2}(P) = 0 \\ \infty, & \text{if } Freq_{\mathcal{C}_1}(P) \neq 0 \text{ and } Freq_{\mathcal{C}_2}(P) = 0 \\ \frac{Freq_{\mathcal{C}1}(P)}{Freq_{\mathcal{C}_2}(P)}, & \text{otherwise.} \end{cases} \quad (1)$$

A growth rate greater than 1 in a quality class sub-dataset $\mathcal{C}_1$ with respect to another quality class sub-dataset $\mathcal{C}_2$ means that the partially ordered pattern is more frequent in $\mathcal{C}_1$ than in $\mathcal{C}_2$. A closed partially ordered pattern must comply with this condition to be considered as a discriminant closed partially ordered pattern in $\mathcal{C}_1$, i.e. denoted as a DCPO-pattern. In our case, this property is very interesting to extract

Table 6
Sequence cutting example.

Site	Sequence
Site 1	$\langle(\text{PHOS}^{Green})(\text{AZOT}^{Blue})(\text{AZOT}^{Green},\text{PHOS}^{Blue})\boxed{(\text{IBGN}^{Blue})}(\text{AZOT}^{Green},\text{PHOS}^{Yellow})(\text{AZOT}^{Green},\text{PHOS}^{Blue})(\text{AZOT}^{Green})\boxed{(\text{IBGN}^{Yellow})}(\text{PHOS}^{Blue})\rangle$
Site 2	$\langle(\text{AZOT}^{Yellow})(\text{PHOS}^{Orange})(\text{AZOT}^{Orange},\text{PHOS}^{Yellow})(\text{AZOT}^{Green})\boxed{(\text{IBGN}^{Orange})}(\text{AZOT}^{Green},\text{PHOS}^{Yellow})(\text{PHOS}^{Yellow})(\text{AZOT}^{Green},\text{PHOS}^{Green})(\text{AZOT}^{Blue},\text{PHOS}^{Green})\boxed{(\text{IBGN}^{Yellow})}\rangle$

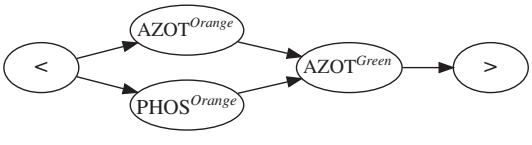


Fig. 6. A partially ordered pattern example.

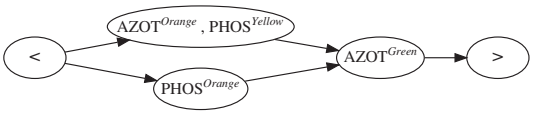


Fig. 7. A closed partially ordered pattern example which brings exactly the same information as the partially ordered pattern in Fig. 6.

environmental markers for each biological quality class sub-dataset. Nevertheless, the growth rate measure defined in the literature is not adapted for a number of quality class sub-datasets higher than two. It is only possible to apply the measure on a quality class sub-dataset with respect to another. We then propose a generalized growth rate that considers multiple quality class sub-datasets (Definition 2).

Definition 2. *Given a closed partially ordered pattern P, a quality class sub-dataset $\mathcal{C}$ and a set of quality class sub-datasets $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n\}$, the generalized growth rate of P in $\mathcal{C}$ with respect to $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n\}$, denoted $GGR(P, \mathcal{C}\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n\})$, is defined as*

$$\begin{cases} 0, Freq_{\mathcal{C}}(P) = 0 \text{ and } max\left(Freq_{\mathcal{C}1}(P), Freq_{\mathcal{C}2}(P), \ldots, Freq_{\mathcal{C}_n}(P)\right) = 0 \\ \infty, \text{ if } Freq_{\mathcal{C}}(P) \neq 0 \text{ and } max\left(Freq_{\mathcal{C}1}(P), Freq_{\mathcal{C}2}(P), \ldots, Freq_{\mathcal{C}_n}(P)\right) = 0 \\ \dfrac{Freq_{\mathcal{C}}(P)}{max\left(Freq_{\mathcal{C}1}(P), Freq_{\mathcal{C}2}(P), \ldots, Freq_{\mathcal{C}_n}(P)\right)}, \text{ otherwise.} \end{cases} \quad (2)$$

This generalized definition leads to compute the growth rate of a pattern in a quality class sub-dataset $\mathcal{C}$ with respect to the maximal frequency in a set of other quality class sub-datasets. Retrieving the maximal frequency of a pattern P in other quality class sub-datasets ensures that with a generalized growth rate greater than 1 in $\mathcal{C}$, the frequency of P in $\mathcal{C}$ is higher than its frequency in all other quality class sub-datasets, thus P is discriminant in $\mathcal{C}$. For example, given a minimum frequency $\theta = 30\%$, the closed partially ordered pattern in Fig. 7 is extracted from quality class sub-datasets IBGN^{Yellow} and IBGN^{Orange} with a frequency of 1/3 and 2/2 respectively. This closed partially ordered pattern is removed from quality class sub-dataset IBGN^{Yellow} since it is more frequent in quality class sub-dataset IBGN^{Orange}. We note that with the generalized growth rate measure, a DCPO-pattern cannot be discriminant in two quality class sub-datasets simultaneously.

In several pattern mining applications, the amount of extracted patterns may be huge with thousand or millions of extracted patterns.

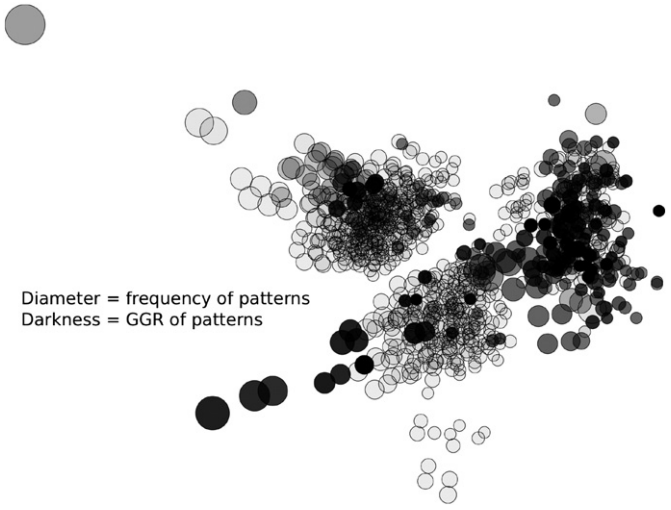


Fig. 8. DCPO-pattern distribution from quality class sub-dataset IBGN^{Red}.

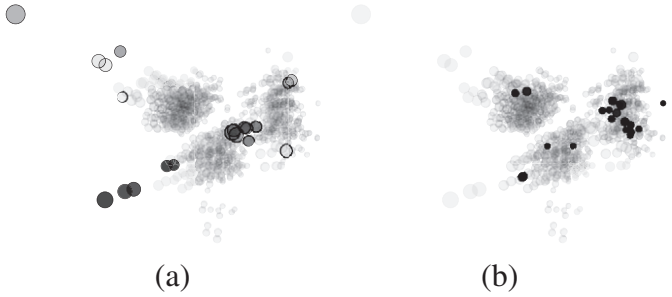


Fig. 9. Preliminary tests: the 20 most frequent (a) and 20 most discriminant (b) DCPO-patterns.

Analyzing such results is extremely complex and requires the use of interestingness measures (Geng and Hamilton, 2006) to filter and reduce the result set. Section 3.3 introduces a selection operation based on some identified dimensions for analysis.

3.3. Interestingness measures and pattern selection

To solve the problem of the huge number of patterns, we define a method to filter the k most interesting DCPO-patterns for analysis. We then define the notion of interestingness specific to this work. After some preliminary tests and discussions, we have identified three different relevant aspects on DCPO-patterns:

1. *Frequency*: represented by the number of sequences in quality class sub-datasets that support the DCPO-patterns (Section 3.2). Analysts are interested in the most frequent DCPO-patterns per quality class sub-dataset since high frequency DCPO-patterns are more supported in the dataset than low frequency DCPO-patterns.
2. *Discriminance*: represented by the generalized growth rate value (Section 3.2). Analysts are interested in the most discriminant DCPO-patterns per quality class sub-dataset. The more discriminant a DCPO-pattern is in a quality class sub-dataset, the more specific it is in this quality class sub-dataset.
3. *Redundancy*: in DCPO-pattern mining, it is usual to find DCPO-patterns that carry almost the same knowledge and that describe almost the same set of sequences. Avoiding redundancy in results is important to improve DCPO-pattern analysis, by retrieving an example of each different case found in the data, without retrieving similar cases.

The ideal result is a small set of DCPO-patterns that are the most frequent and the most discriminant, while they retrieve a viewpoint on the diversity in the data. Nevertheless, selecting such a small subset in thousands of DCPO-patterns is not an easy task. We illustrate this issue by projecting information about extracted DCPO-patterns from quality class sub-dataset IBGN^{Red} (Fig. 8). This projection is obtained by performing a multidimensional scaling approach (Brog and Groenen, 1997) that aims at modeling dissimilarity among pairs of objects into

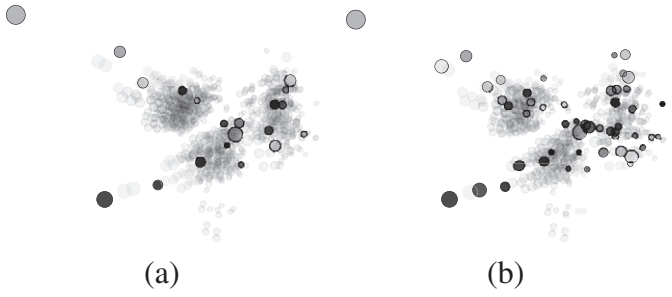


Fig. 10. *PB_Index* tests: the 20 (a) and the 50 (b) most balanced DCPO-patterns from quality class sub-dataset IBGN^{Red}.

Table 8
Biological quality class sub-datasets.

	IBGN	IBD	IPR
Blue	1056	1076	52
Green	2405	1375	162
Yellow	1282	532	126
Orange	556	108	76
Red	89	17	62

points in a low-dimensional geometric space. These points can be represented graphically and visualized afterwards. For our purpose, points represent DCPO-patterns and the dissimilarity measure is the Hamming distance (Hamming, 1950) (explained below).

In this two dimension space, each circle represents a discriminant DCPO-pattern: its diameter is proportional to the frequency of the DCPO-pattern while its darkness is proportional to the discriminance (generalized growth rate value). Two circles far from each other mean that the corresponding DCPO-patterns are supported by different sequences of the dataset, thus they probably contain a different information. The Hamming distance is used to compute the distance between two DCPO-patterns according to the set of sequences supporting them. For example, a dataset of six sequences can be represented with a binary code. A DCPO-pattern with the binary code 011101 means that it is supported by all sequences of the dataset except the first and the fifth sequences (1 when the DCPO-pattern is supported, 0 otherwise). The Hamming distance between two binary codes is the number of bits that differ. Then, the Hamming distance between binary codes 010101 and 011100 is 2 because they differ at the third and the sixth indices. In the projection, we observe that the most frequent DCPO-patterns are rarely the most discriminant and vice versa. Furthermore, many DCPO-patterns are grouped into clusters (almost three in the example). It means that in the result set, a lot of DCPO-patterns are mostly supported by the same sequences of the dataset.

Given this big set of DCPO-patterns, the aim is to select the k top interesting DCPO-patterns. We have first tried two selection approaches based on the k most frequent and on the k most discriminant DCPO-patterns. Based on the extracted DCPO-patterns from quality class sub-dataset $IBGN^{Red}$ and their projection in Fig. 8, we selected and projected the 20 top DCPO-patterns according to their frequency (Fig. 9a) and the 20 top DCPO-patterns according to their discriminance (Fig. 9b). Based on Fig. 8, all the unselected DCPO-patterns in Fig. 9a and b are grayed.

Selecting the most frequent DCPO-patterns leads to select the most common DCPO-patterns in the dataset. We observe that DCPO-patterns contained in the dense clusters are not selected. DCPO-patterns in clusters are in majority less frequent but their darkness means that they have a high generalized growth rate and are very discriminant. Conversely, selecting the most discriminant DCPO-patterns leads to only select DCPO-patterns in clusters and skipping some more frequent DCPO-patterns that probably better describe the dataset. Both approaches present the same drawback. Many selected DCPO-patterns are often very closed in the projection. Thus they probably carry a very similar information and are redundant, which is not relevant for analysis.

3.4. Combination of the dimensions of analysis

To tackle this issue, we choose to combine the distance, the generalized growth rate and the frequency. Given the set $\mathcal{P}$ of unselected DCPO-patterns and the set $\mathcal{SP}$ of already selected DCPO-patterns, the aim is to compute a score of interestingness between 0 and 1 for each $P \in \mathcal{P}$. Then, the DCPO-pattern P with the highest score is added to the set of selected DCPO-patterns $\mathcal{SP}$. To give exactly the same weight to the three dimensions, we have normalized the generalized growth rate measure and the Hamming distance to calculate a value between 0 and 1 for the two measures. The frequency is already a score between 0 and 1, i.e. a frequency equal to 0.2 corresponds to a DCPO-pattern supported by 20% of the dataset.

Our balanced measure is denoted *PB_Index* for Pattern Balance Index. Given the set $\mathcal{P}$ of unselected DCPO-patterns and the set $\mathcal{SP}$ of already selected DCPO-patterns, for each $P \in \mathcal{P}$ the measure is defined as follows:

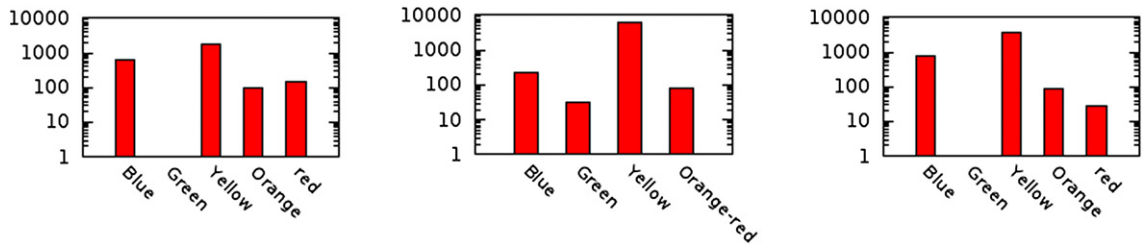
$$PB\_Index(P, \mathcal{SP}) = \frac{NMH(P, \mathcal{SP}) \times Freq(P) \times NGGR(P)}{NMH(P, \mathcal{SP}) + Freq(P) + NGGR(P)} \times 3 \tag{3}$$

(4)

with $NMH(P, \mathcal{SP})$ the minimal normalized Hamming distance between P and the set of patterns $\mathcal{SP}$ (this distance is equal to 1 when the set of selected DCPO-patterns is empty), $Freq(P)$ the frequency of P and $NGGR(P)$ the normalized generalized growth rate of P in the considered quality class sub-dataset. In the numerator, the multiplication between each dimension of analysis allows us to penalize DCPO-patterns for which the value for one or more dimensions is low, thus a selected DCPO-pattern is balanced between dimensions, i.e. each dimension is considered as important. The multiplicative factor 3 is for normalization, i.e. obtaining a score between 0 and 1. Given a parameter k, this method adds iteratively the most balanced DCPO-pattern to the set of selected DCPO-patterns until the size of the set is equal to k. Fig. 10a and b shows experiments performed on DCPO-patterns extracted from quality class sub-dataset $IBGN^{Red}$ by selecting the 50 and the 20 most balanced DCPO-patterns.

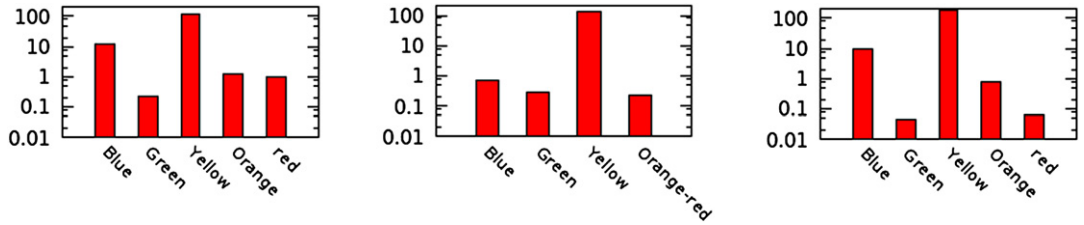
We observe that the selected DCPO-patterns are more diverse (less redundant) than the k most frequent or the k most discriminant, while they are a good balance between the most frequent and the most discriminant. In addition, increasing the parameter k leads to explore even more the diversity of the overall DCPO-patterns set, with a minimized redundancy.

We now present the obtained results.



(a) IBGN (b) IBD (c) IPR

Fig. 11. Number of DCPO-patterns for each biological quality value.



(a) IBGN (b) IBD (c) IPR

Fig. 12. Computation time (in seconds) of DCPO-patterns for each biological quality value.

4. Results

In this section, we provide experiments for three bio-indices available in the Fresqueau dataset and discussed in Section 3.1. As a reminder, these bio-indices are IBGN, IBD and IPR that correspond to macroinvertebrate, diatom and fish dimensions, respectively. By applying the process on each index, we generate their corresponding biological datasets. According to expert knowledge, for each bio-index we have chosen different time intervals in order to capture sequences that precede a biological sampling. Thus, we select a time interval corresponding to four months before a measurement of IBGN and IPR, and we select a time interval of two months before a measure of IBD. The reason is that diatom populations renew faster than macro-invertebrates and fishes. Then, it is relevant to restrict the interval of physico-chemical sequences on the IBD bio-index. Table 8 synthesizes the number of sequences obtained for each quality class sub-dataset. For instance, we generated 1375 sequences for the quality class sub-dataset IBD^{Green}. We obtain very unbalanced quality class sub-datasets: the green IBGN quality class sub-dataset contains 2405 sequences while the red IBGN quality class sub-dataset contains 89 sequences.

The red IBD quality class sub-dataset is problematic with only 17 sequences. Then, with a minimum frequency threshold of 10%, DCPO-patterns supported by only two sequences are extracted. In statistics and data mining domains, extracting knowledge from two data observations is not relevant. Thus, we merged red and orange IBD quality class sub-datasets in a new quality class sub-dataset $IBD^{Orange\text{-}Red}$ composed of 125 sequences. We now show results obtained from each index.

4.1. Overall results

We mined DCPO-patterns with a minimum frequency threshold θ of 10% in all biological datasets. We choose this value because it allows us to extract a significant number of DCPO-patterns from the data. Extracting below this minimum frequency threshold is less statistically significant, especially for sub-datasets with a small number of sequences.

Fig. 11a, b and c shows the number of DCPO-patterns extracted in each quality class sub-dataset. We observe that the number of DCPO-patterns per quality class sub-dataset is unbalanced. In order to highlight this aspect, a logarithmic scale is used on the ordinate axis. For example, 6202 DCPO-patterns are extracted from quality class sub-dataset IBD^{Yellow} while 31 DCPO-patterns are extracted from quality class sub-dataset IBD^{Green}. For the three bio-indices, the yellow quality class sub-dataset is always the one having the biggest amount of DCPO-patterns, at least two times over other quality class sub-datasets. Furthermore, no DCPO-patterns are found from the IPR green quality class sub-dataset and only one DCPO-pattern is extracted from the IBGN green quality class sub-dataset.

The computation time for each quality class sub-dataset is given by Fig. 12a, b and c. We observe that the number of DCPO-patterns extracted is very correlated. Indeed, pattern mining approaches are almost linear according to the number of extracted patterns. The maximal computation time is 177.11 s for the quality class sub-dataset IPR^{Yellow} (3554 patterns) and the minimal one is 0.046 s for the quality class sub-dataset IPR^{Green} (0 pattern). Thus, the computation time is acceptable since this hydrobiological application does not require a real time approach. It does not exceed 3 min in the worst case.

In the next section are presented more detailed results for each bio-index using the selection operation presented in Section 3.3.

4.2. Filtered patterns

For each biological dataset (IBGN, IBD and IPR), we present two DCPO-patterns extracted from the red and blue quality class sub-datasets, i.e. the two extreme quality classes, except for IBD where the orange and the red quality class sub-datasets are merged. These DCPO-patterns are picked from the set of selected DCPO-patterns obtained with the *PB_Index* measure with a parameter $k = 15$, i.e. we extracted the 15 most balanced DCPO-patterns. We choose the number of 15 DCPO-patterns empirically since after many tests, we observe that it provides a small set of results easy to analyze and it captures the diversity contained in the data. The set of the 15 most balanced DCPO-patterns per bio-index quality class is accessible at this webpage[3] (in color). For the sake of clarity, DCPO-patterns presented in this paper contain a maximum of four items and two timestamps, but more complex DCPO-patterns are provided on the webpage.

4.2.1. IPR

By analyzing fish taxa in a river, this bio-index aims at providing an evaluation of the river quality. For a river station, it measures the gap between the current fish population and the fish population that should be present without human activity impact. Experts also consider parameters such as the watershed surface, the average air temperature and the width and the depth of the river station. An IPR score of 0 means that there is no difference between the measured situation and the ideal situation. As this score increases, the gap between the current situation and the ideal situation is more important. Figs. 13 and 14 provide two DCPO-patterns extracted from blue IPR quality class sub-dataset and red IPR quality class sub-dataset, respectively. The DCPO-pattern in Fig. 13 has a frequency equal to 39.62% in the blue IPR quality class sub-dataset and 16.12% in the red one.

4.2.2. IBGN

This biological dimension has already been briefly presented in Section 3.1. It is based on the presence or absence of over 100 pollusensitive macro-invertebrate taxa. The obtained value (a score between 0 and 20) is based on macro-invertebrate taxa and on their abundance in sampling. A score of 20 is representative of a very good river quality, while a score of 0 is characteristic of a very bad river quality. As for IBGN bio-index, we present two extracted DCPO-patterns in Figs. 15 and 16.

4.2.3. IBD

This last bio-index concerns the microscopic granularity of the biology. Like IPR and IBGN bio-indices, IBD is the measurement of a score that highlights the viability of the river. Here, the viability is

[3] http://engees-fresqueau.unistra.fr/patterns/patterns.html.

Fig. 13. DCPO-pattern from quality class sub-dataset IPR^{Blue} with frequencies: Blue = 39.62%, Green = 25.34%, Yellow = 17.07%, Orange = 20.28%, Red = 16.12%.

Fig. 14. DCPO-pattern from quality class sub-dataset IPR^{Red} with frequencies: Blue = 5.66%, Green = 7.53%, Yellow = 5.69%, Orange = 4.34%, Red = 9.67%.

representative of micro-algae populations. It consists in analyzing and counting 400 individuals from a sample. As with IBGN, the calculation is based on a data dictionary providing the polluo-sensitivity and the ability of the species to be in various environments. Figs. 17 and 18 show two examples of DCPO-patterns on this bio-index.

The above experiments are performed to show the methodology and some obtained results. We now present a qualitative study of the proposed process.

5. Discussion

This section is a discussion about the advantages of mining DCPO-patterns over other data mining techniques. For the clarity reasons, Table 9 reminds the relationship between the colors and the quality categories.

5.1. DCPO-patterns meet with domain knowledge

Here, we discuss about the fact that DCPO-patterns well match with existing knowledge in hydrobiology, and also provide new knowledge. We take as example IBGN and IPR indices.

IBGN index has been mainly developed to address the problem of organic matter pollutions (MOOX) (Vernaux et al., 1982). With DCPO-patterns, we are able to retrieve this knowledge: blue MOOX appears in blue IBGN patterns (Fig. 15) while a red MOOX appears in red IBGN patterns (Fig. 16). Thus, in the case of the IBGN index, the analysis of patterns shows a correspondence between MOOX quality classes and IBGN quality classes. For example, a blue MOOX is observed with a blue IBGN, a yellow MOOX is observed with a yellow IBGN, etc. However, the impact of phosphorus pollutions (macro-parameter PHOS) on the IBGN index value is a lesser-known fact and DCPO-patterns also show a correspondence between PHOS quality classes and IBGN. Such a correspondence is interesting since it may highlight that the source of phosphorus pollutions has an impact on macro-invertebrates (IBGN index).

Concerning the IPR index, the DCPO-pattern in Fig. 13 shows that very good fish communities need very good quality river. Indeed, we know that fishes are sensitive to organic and nitrogenous matters (because of oxygen consumption during biodegradation of organic matter). On the other hand (Fig. 14), a bad fish community can exist with no bad physico-chemical parameters if hydromorphological conditions are not sufficient (concrete channels, dam, etc.). Indeed, hydromorphological conditions have an important impact on fish habitat.

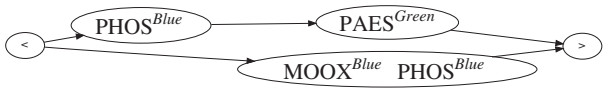


Fig. 15. DCPO-pattern from quality class sub-dataset $IBGN^{Blue}$ with frequencies: Blue = 16.84%, Green = 10.12%, Yellow = 5.41%, Orange = 1.77%, Red = 0%.

Fig. 16. DCPO-pattern from quality class sub-dataset $IBGN^{Red}$ with frequencies: Blue = 0%, Green = 0.16%, Yellow = 1.23%, Orange = 7.8%, Red = 19.1%.

Fig. 17. DCPO-pattern from quality class sub-dataset IBD^{Blue} with frequencies: Blue = 14.95%, Green = 4.58%, Yellow = 1.21%, Orange-Red = 0.74%.

5.2. Using domain knowledge to discretize the data

Discretizing data with domain knowledge (SEQ-eau and AFNOR standards) is useful. By extracting DCPO-patterns by quality class sub-datasets (biological quality classes), we can highlight and measure the concordance between physico-chemical and biological parameter quality classes. For example, the impact of the physico-chemical parameter PHOS on diatoms has been widely studied (Coring, 1999; Kelly et al., 1995) and is well-known. The river biological quality class for diatoms is often bad when very high PHOS concentrations are measured. Currently, it is not possible to predict precisely its impact on IBD index value. It is just possible to say that a moderate/bad biological quality class, e.g. a yellow, an orange or a red IBD class, is predictable with high probability. By analyzing some DCPO-patterns, we observe that for each IBD quality class sub-dataset, a PHOS quality class is related. Table 10 sums it up: we can observe that the IBD index quality classes do not correspond exactly with physico-chemical quality classes. IBD quality classes decrease more rapidly than PHOS quality classes. For instance, a yellow IBD is related to a green PHOS and an orange IBD is related to a yellow PHOS. Furthermore, a red IBD is related to a red PHOS, but specifically concerned by a red PHOS combined with a red AZOT and a red MOOX. Then, with IBD, we observe that: (1) the biological status is very sensitive to water disturbances and (2) the combination of multiple bad physico-chemical values produces necessarily degradation of the biological dimension. It leads to new perspectives: suggesting to determine new interval values for physico-chemical parameter quality classes to match with bio-index classes and to evaluate more precisely the impact of the combination of physico-chemical parameters.

Table 9
Correspondence between colors and quality categories.

Color	Quality category
Blue	Very good
Green	Good
Yellow	Medium
Orange	Bad
Red	Very bad

Table 10
Correspondence between IBD quality classes and PHOS quality classes.

IBD	Physico-chemistry
Yellow	$PHOS^{Green}$
Orange	$PHOS^{Yellow}$
Red	$PHOS^{Red}$
100% Red	$PHOS^{Red}$, $AZOT^{Red}$, $MOOX^{Red}$

Fig. 18. DCPO-pattern from quality class sub-dataset IBD^{Red} with frequencies: Blue = 0%, Green = 0.12%, Yellow = 1.03%, Orange-Red = 12.68%.

Furthermore, when analyzing the physico-chemistry, sampled values are usually aggregated by using percentiles or a statistical average. The interpretation of results is thus sensitive to the aggregation method. Using temporal patterns such as DCPO-patterns avoids the aggregation issue since it is based on sequences of discretized parameters. Thus, all sampled parameters are taken into account.

5.3. Selection of the most balanced DCPO-patterns

The Pattern Balance Index proposed in Section 3.3 highlights with a few dozen of DCPO-patterns the knowledge diversity contained in several hundreds of thousands DCPO-patterns. It allows hydrobiologists to mine the data at low minimum frequency thresholds to discover and capture less frequent but new and potentially interesting knowledge without having to manage and analyze a huge volume of DCPO-patterns. It provides a global overview of the results.

5.4. Benefits of applying DCPO-pattern approaches

This paper presents the application of a DCPO-pattern method to discover links between physico-chemistry values and biological values. Furthermore, this approach could be useful for many other hydrobiological or environmental problems.

On such temporal data, sequential pattern approaches are also a possible alternative to DCPO-patterns. Compared to DCPO-patterns, sequential patterns are represented by a flat list of symbolic values. DCPO-patterns have the advantage of taking into consideration elements at different timestamps with no temporal link. For example, let us look at the DCPO-pattern provided by Fig. 17, $AZOT^{Blue}$ and $PHOS^{Blue}$ are frequently measured at the same time and, in parallel, a $MOOX^{Green}$ is measured without being ordered. A sequential pattern mining approach would led to extract two different sequential patterns: $\langle(AZOT^{Blue}, PHOS^{Blue})\rangle$ and $\langle(MOOX^{Green})\rangle$. Such a division of the knowledge with sequential patterns is a drawback because the information about the co-existence of physico-chemical parameters, even if they are not temporally ordered, allows hydrobiologists to better identify which parameters and which combinations have a stronger impact. In our work, this feature is very important because hydrobiologists make the assumption that many physico-chemical parameters are not temporally correlated between them.

Thus DCPO-patterns are well-suited for the problem of identifying temporal observations with many variables. Indeed, such patterns are able to capture groups of variables that are observed at the same time, at the following timestamp or at different timestamps with no temporal link. Each DCPO-pattern corresponds to a frequent information in the data and it does not provide any information about the correlation between variables. Nevertheless, it could be extended to obtain such correlations by post-processing DCPO-patterns with approaches based on association rules (Agrawal and Srikant, 1994) or sequential rules (Fournier-Viger et al., 2011).

This technique and more generally temporal pattern approaches are sensitive to the discretization process. However, this process is also well-adapted when there exists domain knowledge. Indeed, continuous values are not always easy to analyze and discretized values can provide simple and robust information based on previous works, such as SEQ-eau and AFNOR standards.

Pattern mining approaches are adapted to the case of hundred or thousand instances in the data. It is possible to first use such methods to obtain an overview of the knowledge contained in the data, and then to process multivariate statistics (Legendre and Legendre, 2012). The idea is to refine the knowledge by specifically analyzing parameters frequently observed in patterns. Thus, pattern mining methods could be a good addition to statistical approaches.

5.5. Perspectives

We applied this generic process on the three bio-indices IBGN, IBD and IPR because they are studied and measured for many years in French river ecosystems. It then gives us a substantial amount of data to explore. However there exist more recent bio-indices that concern the river viability for macrophytes with the IBMR bio-index (AFNOR (Association Française de NORmalisation), 2003) and oligochetes with the IOBS bio-index (AFNOR (Association Française de NORmalisation), 2002). We are interested in analyzing them to discover new knowledge on rivers. Exploring all the biological dimensions is important to measure a global river quality.

In addition, since DCPO-patterns are useful to capture the discriminant features for different biological qualities, we wish to extend our process to classification (Cheng et al., 2007, 2008) and prediction (Wang et al., 2008) of river quality. This perspective completely matches with the objectives of the European Water Framework Directive. Measuring bio-indices on the field leads to long delays in analysis. Predicting accurately the river viability for biological dimensions by just analyzing recent physico-chemical samplings could then be straightforward. Many methods mainly focus on extreme biological quality classes (blue and red), but we show that taking into account intermediate quality classes is mandatory to improve the understanding of river ecosystems.

6. Conclusion

This article proposes a temporal pattern based approach applied to hydro-ecological data to extract the relations between the physico-chemistry and the biology.

The proposed method is generic since it takes into consideration the different biological dimensions of river ecosystems. It highlights the link of physico-chemistry with biology and points the way towards new perspectives on the quantification of these relations.

In future work, we aim to use the extracted discriminant closed partially ordered patterns as physico-chemical bio-markers to perform the classification, or the prediction of river quality according to the biology to reach good quality in rivers as required by the European Water Framework Directive.

Acknowledgment

This work was funded by the French National Research Agency (ANR), as part of the ANR_{11}_$MONU_{14}$ Fresqueau project.

References

AFNOR (Association Française de NORmalisation), 1992, révision 2004. Qualité de l'eau: détermination de l'Indice Biologique Global Normalisé (IBGN). Norne Française NF T90-350.
AFNOR (Association Française de NORmalisation), 2000, révision 2007. Qualité de l'eau: détermination de l'Indice Biologique Diatomées (IBD). Norne Française NF T90-354.
AFNOR (Association Française de NORmalisation), 2002. Qualité de l'eau: détermination de l'Indice Oligochètes de Bioindication des Sédiments (IOBS). Norne Française NF T90-390.
AFNOR (Association Française de NORmalisation), 2003. Qualité de l'eau: détermination de l'Indice Biologique Macrophytique en Rivière (IBMR). Norne Française NF T90-395.
AFNOR (Association Française de NORmalisation), 2004. Qualité de l'eau: détermination de l'Indice poissons rivière (IPR). Norne Française NF T90-344.
Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules in large databases. International Conference on Very Large Data Bases, VLDB, pp. 487–499.

Agrawal, R., Srikant, R., 1995. Mining sequential patterns. International Conference on Data Engineering, ICDE, pp. 3–14.
Bertaux, A., Le Ber, F., Braud, A., Trémolières, M., 2009. Identifying ecological traits: a concrete FCA-based approach. Formal Concept Analysis vol. 5548, pp. 224–236.
Brog, I., Groenen, P.J.F., 1997. Modern Multidimensional Scaling: Theory and Applications. Springer-Verlag.
Cheng, H., Yan, X., Han, J., Hsu, C., 2007. Discriminative frequent pattern analysis for effective classification. International Conference on Data Engineering, ICDE, pp. 716–725.
Cheng, H., Yan, X., Han, J., Yu, P.S., 2008. Direct discriminative pattern mining for effective classification. International Conference on Data Engineering, ICDE, pp. 169–178.
Coring, E., 1999. Situation and developments of algal (diatom)-based techniques for monitoring rivers in Germany. Use of Algae for Monitoring Rivers III, pp. 122–127.
D'heygere, T., Goethals, P.L., Pauw, N.D., 2003. Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. Ecol. Model. 160, 291–300.
Dakou, E., D'heygere, T., Dedecker, A., Goethals, P., Lazaridou-Dimitriadou, M., Pauw, N., 2007. Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece). Aquat. Ecol. 41, 399–411.
Dedecker, A.P., Goethals, P.L., Gabriels, W., Pauw, N.D., 2004. Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrates in the Zwalm river basin (Flanders, Belgium). Ecol. Model. 174, 161–173.
Dong, G., Li, J., 1999. Efficient mining of emerging patterns: discovering trends and differences. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD, pp. 43–52.
E. Union, 2000. 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. Off. J. OJ L 327, 1–73.
Fabrègue, M., Braud, A., Bringay, S., Ber, F., Teisseire, M., 2013. OrderSpan: mining closed partially ordered patterns. Advances in Intelligent Data Analysis XII vol. 8207, pp. 186–197.
Fournier-Viger, P., Nkambou, R., Tseng, V.S.-M., 2011. Rulegrowth: mining sequential rules common to several sequences by pattern-growth. Proceedings of the 2011 ACM Symposium on Applied Computing, pp. 956–961.
Geng, L., Hamilton, H.J., 2006. Interestingness measures for data mining: A survey. ACM Computing Survey 38.
George, A., Binu, D., 2012. DRL-Prefixspan: a novel pattern growth algorithm for discovering downturn, revision and launch (DRL) sequential patterns. Cent. Eur. J. Comput. Sci. 2, 426–439.
Goethals, P.L., Dedecker, A., Gabriels, W., Lek, S., Pauw, N., 2007. Applications of artificial neural networks predicting macroinvertebrates in freshwaters. Aquat. Ecol. 41, 491–508.
Hamming, R.W., 1950. Error detecting and error correcting codes. Bell Syst. Tech. J. 29, 147–160.
Kelly, M.G., Penny, C.J., Whitton, B.A., 1995. Comparative performance of bentic diatom indices to asses river water quality. Hydrobiologia 302, 179–188.
Kocev, D., Naumoski, A., Mitreski, K., Krstić, S., Džeroski, S., 2010. Learning habitat models for the diatom community in Lake Prespa. Ecol. Model. 221 (2), 330–337.
Legendre, P., Legendre, L.F., 2012. Numerical ecology vol. 20. Elsevier.
Recknagel, F., Ostrovsky, I., Cao, H., Zohary, T., Zhang, X., 2013. Ecological relationships, thresholds and time-lags determining phytoplankton community dynamics of Lake Kinneret, Israel elucidated by evolutionary computation and wavelets. Ecol. Model. 6, 1–3.
Ren, J., Wang, L., Dong, J., Hu, C., Wang, K., 2009. A novel sequential pattern mining algorithm for the feature discovery of software fault. International Conference on Computational Intelligence and Software Engineering, Vol. 5854 of CiSE, pp. 439–447.
Sallaberry, A., Pecheur, N., Bringay, S., Roche, M., Teisseire, M., 2011. Sequential patterns mining and gene sequence visualization to discover novelty from microarray data. J. Biomed. Inform. 44, 760–774.
Van Dam, H., Mertens, A., Sinkeldam, J., 1994. A coded checklist and ecological indicators values of freshwater diatoms from the Netherlands. Neth. J. Aquat. Ecol. 117–134.
Vernaux, J., Galmiche, P., Janier, F., Monnot, A., 1982. Une nouvelle méthode pratique d'évaluation de la qualité des eaux courantes. Un indice biologique de qualité générale (IBG). Ann. Sci. Univ. Franche-Comté Besançon 11–21.
Wang, M., Shang, X., Li, Z., 2008. Sequential pattern mining for protein function prediction. Advanced Data Mining and Applications, Vol. 5139 of ADMA, pp. 652–658.
Yang, Y.C.E., Cai, X., Herricks, E.E., 2008. Identification of hydrologic indicators related to fish diversity and abundance: A data mining approach for fish community analysis. Water Resources Research 44.