A knowledge discovery process for spatiotemporal data: Application to river water quality monitoring

Hugo Alatrista Salas, Jérôme Azé, Sandra Bringay, Flavie Cernesson, Nazha Selmaoui-Folcher, Maguelonne Teisseire

A Knowledge Discovery Process for Spatiotemporal Data: Application to River Water Quality Monitoring

H. Alatrista-Salas[a,c,*], J. Azé[b], S. Bringay[b], F. Cernesson[d], N. Selmaoui-Folcher[c], M. Teisseire[a]

[a]*Irstea-TETIS, 500, rue J. F. Breton 34093, Montpellier Cedex 5, France*
[b]*LIRMM, 161, rue Ada 34392, Montpellier Cedex 5, France*
[c]*PPME - BP R4 98851, Nouméa Cedex, New Caledonia*
[d]*AgroParisTech-TETIS, 500 rue J. F. Breton 34093, Montpellier Cedex 5, France*

Abstract

Rapid population growth, and human activities (such as agriculture, industry, transports,...) development have increased vulnerability risk for water resources. Due to the complexity of natural processes and the numerous interactions between hydro-systems and human pressures, water quality is difficult to be quantified. In this context, we present a knowledge discovery process applied to hydrological data. To achieve this objective, we combine successive methods to extract knowledge on data collected at stations located along several rivers. Firstly, data is pre-processed in order to obtain different spatial proximities. Later, we apply a standard algorithm to extract sequential patterns. Finally we propose a combination of two techniques (1) to filter patterns based on interest measure, and; (2) to group and present them graphically, to help the experts. Such elements can be used to assess spatialized indicators to assist the interpretation of ecological and river monitoring pressure data.

Keywords:
Data mining, Spatiotemporal databases, Sequential patterns, Water management

*Corresponding author
Email addresses: hugo.alatrista-salas@teledetection.fr (H. Alatrista-Salas), jerome.aze@lirmm.fr (J. Azé), sandra.bringay@lirmm.fr (S. Bringay), flavie.cernesson@teledetection.fr (F. Cernesson), nazha.selmaoui@univ-nc.nc (N. Selmaoui-Folcher), maguelonne.teisseire@teledetection.fr (M. Teisseire)

1. Introduction

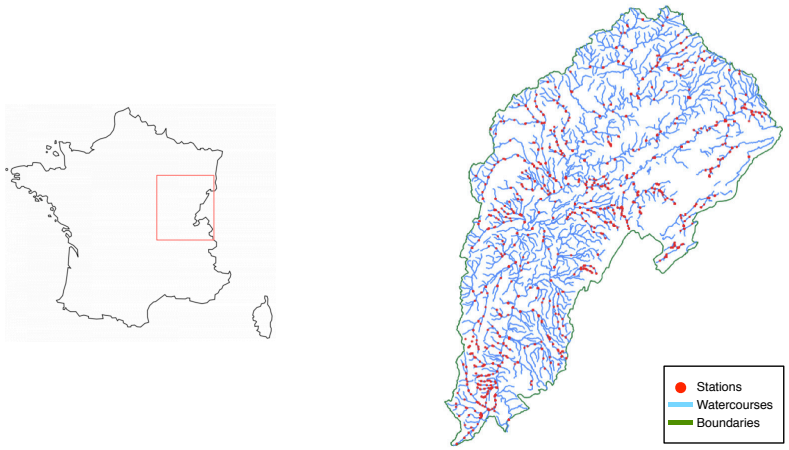
Improvements in digital data collection devices and data storage technology have allowed companies and organizations to store increasingly huge amounts of data thus making it harder to analyze them manually. Therefore, new techniques have been developed to help humans to automatically turn this huge volume of data into useful knowledge that enables a better understanding of phenomena occurring in their environment. These techniques make up *Knowledge Discovery in Databases (KDD)* which is characterized as a multi-step process for discovering valid, novel and potentially useful information.

Natural phenomena involve both spatial and temporal components. For example, in environmental contexts, river pollution is a phenomenon which is observed by measuring physicochemical and biological indicators for water quality. These indicators which evolve over time, depend explicitly on the location of sampling stations strategically located along several rivers.

If systems dedicated to water quality monitoring have existed for several decades, the challenge is now to define indicators to take into account the impact of uses and water quality restoration measures. In this context, to build an efficient tool, spatial relations both metric (e.g., distance) and non-metric (e.g., topology, locations,...) and temporal relations (e.g., before or after) must be considered in the KDD process in order to better understand spatiotemporal phenomena.

In this paper, our objective is to analyze the water quality in the hydrological network of Saône watershed (located in the East of France, see Figure 1). To achieve this goal, we describe a KDD process for hydrological data consisting of: (1) a pre-processing step to transform data by grouping stations that consider their different spatial proximity according to their distance, to membership in a common area,...; (2) a second step dedicated to the extraction of sequential patterns in order to take into account the temporal aspect, and; (3) post-processing step, combining a new interest measure called *the least temporal contradiction* in order to filter sequences to retain only the least contradicted over time. This technique is coupled with another one that allows us to determine the degree of similarity between patterns obtained and regroups them.

This paper is organized as follows: in Section 2, we present a brief



(a) (b)

Figure 1: The Saône river watershed: location (a) and hydrographic network (b)

overview of knowledge discovery process in spatiotemporal data. After, in Section 3, we describe a framework for extracting knowledge. The experiments performed are described in Section 4. Finally, we show the results of our proposals by highlighting the short and medium term perspectives in Section 5.

2. Related work

Knowledge discovery in databases (*KDD*) is a dynamic research field. Fayyad et al. [1] presented the most widely used KDD framework and provides a broad overview of knowledge discovery techniques. Here KDD, was described as a set of interactive and iterative steps: data selection, preprocessing, transformation, data mining, and post processing or interpretation. As mentioned by Fayyad et al. [1], *the basic problem addressed by the KDD process is one of mapping low-level data into other forms that might be more compact, more abstract, or more useful.* Data mining is only a step of this general process. Indeed, using only a data mining technique can lead to the discovery of meaningless patterns for experts. Other steps of the KDD process have been added to deal with this problem. All those steps working together and integrating findings into a unified whole produce new knowledge [2].

The advent of GIS (Geographical Information Systems) technology and the availability of large volume of spatiotemporal data has increased the need for effective and efficient methods to extract unknown and unexpected information. Unfortunately, in many situations, a simple data mining method will often be limited in its ability to retrieve informative knowledge from complex spatiotemporal databases [3]. The specificity of environmental data - and in a more general sense spatiotemporal data, w.r.t. classical data - is the significance of spatial and temporal dimensions for the extraction and interpretation process [4]. In this context, authors in [5] highlight the importance of pre and post-processing in a KDD process concerning spatiotemporal data.

Pre-processing and transformation steps (or more simply **pre-processing**) are directly related to the data mining step. These steps have an important impact on mining results. For example in [6], pre-processing is used to integrate spatial information in the data mining step. Spatial data is converted in spatial predicates. Thanks to this transformation, a classical data mining algorithm can be used to extract spatial patterns. Another important criteria when pre-processing spatial data is the granularity chosen when materializing the spatial data as a single table. Indeed, classical data mining algorithms take a simple table as input and does not consider spatial information directly. If the objective is to study changes in data generated by stations, one way of extracting such spatial patterns is to aggregate information for each station in a single row of the input table. In [7, 8], spatial data is mapped to sets of values or sequences of values.

Several pre-processing techniques in spatiotemporal datasets have been discussed in the literature [9, 10, 11]. Each reference has its own focus such as spatial classification, spatial clustering or spatial association rules. To our knowledge, no works have tried to mine sequential patterns at different spatial granularity levels and then combine their results to obtain more informative and general spatial patterns. In fact, the goal of spatial data mining is to discover spatial patterns and to suggest hypotheses about potential generators of this kind of patterns. This task is not straightforward and requires us to challenge the classical KDD process. In this paper we focus on spatial patterns from the perspective of space division using different levels of spatial granularities. This task was performed to deduce more general patterns by averaging attributes of spatial objects grouped into homogeneous areas.

Post-processing is also an important step in the KDD process. Results generated by data mining algorithms are often difficult to interpret by experts. The number of extracted patterns may be too large and some of them

may not be meaningful for experts. Thus, post-processing techniques are needed to (1) filter more relevant patterns and to (2) display solutions in a user-friendly way.

Many measures exist to filter relevant patterns. In the more general context of association rules with the antecedent $\rightarrow$ consequent form, many measures have been proposed to evaluate the relevance of extracted rules, and hence reduce the set of solutions that a human expert can analyze. A comparative study of several relevance measures is presented in [12]. Classical measures for association rules try to evaluate the independence gap between antecedent and consequent. The confidence [13], number of counter-examples associated to rules [14], statistical astonishment [15, 16] are some examples of such measures.

For spatiotemporal data, several measures have also been proposed. Although this topic has received lot of attentions from the research community [17], to our knowledge, no measure has been proposed to evaluate the spatial and temporal prevalence of a pattern in an easily understandable way for experts. Most spatiotemporal measures evaluate only the "spatial support" of each pattern, i.e., the number of times the pattern occurs in different places. In [8], the authors use the classical version of spatial support for the extraction of spatiotemporal patterns. More efficient measures have been proposed to capture the "spatiality" in a data mining process. For instance, in [18], the authors define the *participation index* as a prevalence measure used in the extraction of co-locations. The participation index measure has been modified for different purposes, for example, for the extraction process of cascade spatiotemporal patterns [19] or for the extraction of confident co-location rules [20].

When the temporal prevalence is considered, resulting measures are very difficult to interpret by experts.

In the post-processing step, comparing two patterns by similarity is a fundamental task that has to be defined before one can apply statistical, machine learning, or data mining methods [21]. In order to get a clear view of data, patterns should be classified or clustered so that semantically similar terms are grouped together. Similarity measures such as *Edit* distance [22] and LCS [23] have been proposed in literature and have been applied in many contexts [24, 25, 26]. The measure S^2MP proposed in [27] has been adopted in this work due to its effectiveness and its applicability to sequences of itemsets (see Section 3.3).

In addition, information visualization is an important aspect to help the

expert in the decision-making task. In this context, some techniques targets the visual representation of patterns extracted in order to help experts to better understand and analyze information (for a survey, see, e.g., [28]). In our work, we propose a graphical representation of solutions in groups of similar patterns using the S^2MP measure and k-medoids clustering algorithm. In the next section we will describe our proposition.

3. A framework for mining spatiotemporal data

In this section, we describe the steps of the general process used to extract knowledge in spatiotemporal database.

3.1. General process

Our approach is divided into four steps: (1) spatial decomposition and aggregation; (2) spatially frequent sequential patterns mining; (3) filtering of patterns according to a temporal interest measure, and finally; (4) restitution of solutions in groups. This general process is illustrated in Figure 2.

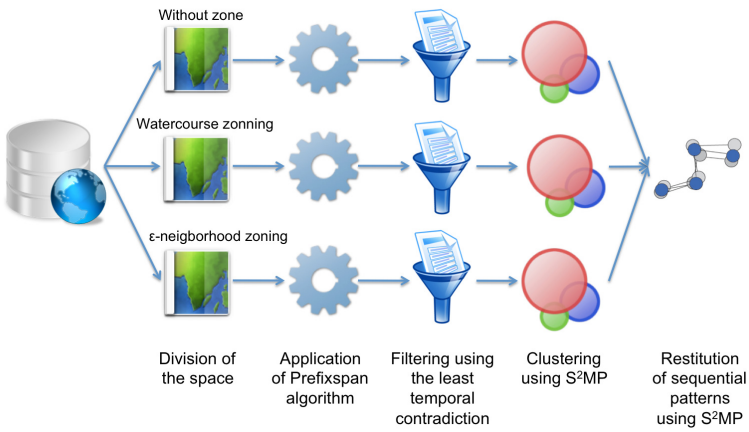


Figure 2: Process of knowledge discovery applied to hydrological data

Spatial decomposition and aggregation are pre-processing steps in which spatial data is mapped to sequences according to different spatial relationships (e.g., station proximity, watercourse). The resulting spatial sequences are the input of the data mining algorithm. The data mining step extracts frequent sequences [13], i.e., those occurring in more than just *minsup* zones (where *minsup* is a user-defined threshold). Therefore, extracted patterns

represent spatially frequent temporal evolutions of zones. The spatial frequency is evaluated by the support measure. Solutions are then filtered according to a new temporal interest measure: *the least temporal contradiction*. This filtering step reduces the number of solutions by eliminating the most contradicted sequences. Later, solutions are grouped using a similarity measure. Finally, all interesting sequences mined at different spatial granularities are combined. Similar sequences are displayed in groups so that experts have access to results which are concise and easier to interpret.

3.2. Spatial pre-processing

Hydrological data are associated to biological indicators collected by monitoring stations strategically positioned along the Saône watershed. This heterogeneous data is also geo-referenced and temporally variable, thus making them difficult to explore globally. Moreover, the spatial relationship between studied objects (i.e., monitoring stations) is implicit. It is therefore necessary to perform pre-processing that takes into account different spatial proximities (e.g., grouping stations according to their distance, according to their association to the same zone,...).

We propose in this step to explore these data in two different ways. The first way is to consider each monitoring station as a unique spatial object so that we can then apply a classical pattern extraction algorithm. The second option is to pre-process data to bring together some monitoring stations and build homogeneous zones of spatial objects. This will enable to study how neighboring station attributes can impact the attributes of the studied station.

For example, in Figure 3, we observe that the monitoring station X can be impacted from both neighbor stations that are located on the same watercourse (represented by yellow lines) or/and monitoring stations that are in a contiguous area but not necessarily positioned on the same watercourse (represented by the red line).

In this context, spatial data can be used to determine the relevant geographical areas to handle (1) flow aspects, by combining the proximity related to watercourse, the flow direction and the connections between the rivers; (2) the spatial proximity of stations, expressed by their Lambert coordinates (geo-referenced coordinate system). Purposely, two spatial divisions were performed in the pre-processing step to divide the space into homogeneous zones:

Figure 3: Impact of neighboring stations to monitoring station X

- A *watercourse* neighborhood approach: for a given watercourse, two stations X and Y located on this watercourse are considered to be neighbors. For example, in Figure 4, stations W, X, Y and Z belong to the same watercourse. These stations are considered to form a single area and their data are combined. An incident that can be studied thanks to this approach is: a fuel outflow from a boat at station X will impact on measures of station X and later on measures on stations Y and Z located on downstream of station X.

- The ϵ-neighborhood: the space is divided into areas adjoining each station by exploiting the Lambert coordinates. In each of these areas, stations that are located within an area of ϵ km^2 centered on station X are grouped, even if these stations belong to different watercourses. For example in Figure 5, stations X and Y are considered to be in the same area, even if they are not on the same watercourse. An example of phenomenon that can be studied due to this approach is: pesticide use in a crop field located between stations X and Y can impact on measures of stations located on rivers around this crop field even if stations are not positioned in the same river.

Thanks to these two spatial division methods, we are able to group the stations within areas and thus to aggregate data in order to extract *spatially frequent sequences*.

In the following sections, we will show that this aggregation provides the most relevant sequential patterns that allows for the heterogeneous nature of

Figure 4: *watercourse* zoning

Figure 5: ϵ-*neighborhood* zoning

the records.

3.3. Sequential patterns mining

Consider the database DB, illustrated in Table 1, which groups all records made by stations along several rivers (e.g., in Table 1, item A could be "good state according bioindicator IBGN value").

Each tuple T is a transaction and consists of a triplet (id-station, date, items): the id of the station, the date of record as well as all current characteristics of the river.

Let $I = \{i_1, i_2, \ldots, i_m\}$ the set of *items* (characteristics). An *itemset* is a non-empty set of items denoted by $(i_1, i_2, \ldots, i_k)$ where i_j is an *item*. A *sequence* S is a non-empty ordered list, of itemsets denoted by $< IS_1, IS_2, \ldots, IS_p >$ where IS_j is an *itemset*.

A *n-sequence* is a sequence of n itemsets.

For example, consider characteristics A, B, C, D and E recorded by the station *Station1* according to the sequence $S =< (A, E)(B, C)(D)(E) >$, as shown in Table 1. This means characteristics A and E were recorded together by *Station1*, i.e., at the same time. Then, *Station1* recorded B and C, the last items in the sequence were recorded later and separately, by the same station. In this example, S is a 4-sequence.

A sequence $< IS_1, IS_2, \ldots, IS_p >$ is a subsequence of another sequence $< IS'_1, IS'_2, \ldots, IS'_m >$ if there exist integers $k_1 < \ldots < k_j < \ldots < k_p$ such as $IS_1 \subseteq IS'_{k_1}, IS_2 \subseteq IS'_{k_2}, \ldots, IS_p \subseteq IS'_{k_p}$. For example, the sequence

Table 1: Example of river characteristics dataset

ID-station	Date	Items
Station1	04/01/12	(A E)
Station2	04/02/28	(E)
Station1	04/03/02	(B C)
Station1	04/03/12	(D)
Station1	04/04/26	(E)

$S' =< (B)(E) >$ is a subsequence of S because $(B) \subseteq (B, C)$ and $(E) \subseteq (E)$. However, $< (B)(C) >$ is not a subsequence of S because the two itemsets (B) and (C) are not included in two different itemsets of S. All characteristics recorded by the same station are grouped and sorted by date. It is called the *data sequence* of the station.

A station *supports* a sequence S if S is included in its data sequence (S is a subsequence of the station data sequence). The *support* of a sequence S is calculated as the percentage of stations that support S.

Let *minsupp* be a minimum support set by the user. A sequence satisfying the minimum support (i.e. whose support is greater or equal than *minsupp*) is a *frequent sequence* called a *sequential pattern*.

The **sequential patterns mining problem** was introduced by [13] in the context of the basket market problem and applied with success in many fields such as biology [29], Web mining [30, 31] or the consumer marketing [32].

As noted above, the main challenge involved in finding sequential patterns in a database is to extract sequences for which support is greater or equal than the specified minimum threshold *minsupp*.

To extract sequential patterns, the PrefixSpan algorithm [33] has been adopted because of its effectiveness with large volumes of data. This method uses a *divide and conquer* strategy by performing a *depth-first search* with successive database projections.

3.4. Filtering with a new interest measure

In the data mining domain, huge number of sequential patterns are frequently obtained. Choosing the most relevant one remains a challenge which is often correlated with the type of data handled.

Even if spatiotemporal data mining has received a lot of attention [34, 35], the study of literature does not report on any work on spatiotemporal interest

measure for sequential patterns. We thus focus on finding sequential patterns that are not contradicted by data over time.

For this, we extend the measure called *the least contradiction (LC)*, defined for association rules in [14, 36] to our context, i.e., sequential patterns.

It must be recalled that this measure, in the context of an association rule $A \rightarrow B$, where A and B are two disjoint sets of items, is defined by:

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

$A\overline{B}$ are itemsets where A is present and B is absent.

We choose to extend the least contradiction measure to sequences of itemsets for two main reasons. First, this measure is simple to understand and to implement. Second, previous work has illustrated the capacity of this measure to extract nuggets of knowledge [14] and to resist to noise [36]. Other measures such as *lift* could also be extended to sequences of itemsets, however, its definition is close to the one of the least contradiction.

Definition: let S be a sequential pattern, the *Least Temporal Contradiction* of S, denoted $LTC(S)$, is defined by:

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

where $\begin{cases} S_d & \text{the set of sequential patterns including all itemsets of the sequence } S \text{ but in a different position} \\ S_t & \text{the set of sequential patterns including all items which appeared in sequence } S \end{cases}$

LTC allows us to keep the original spirit of the least contradiction measure which was designed to estimate how many times a sequential pattern is verified *vs* how many times it is disabled. A sequence that is most frequently tested as enabled is *a priori* relevant. Like the conventional version, the LTC is normalized. Here, normalization is performed in relation to the sum of supports of the sequences that can be built from the items composing the studied sequence.

For example, consider the following sequential patterns and their support:

$$\begin{cases} S_1 =< (AB)(BC) > & , supp(S_1) = 0.25 \\ S_2 =< (BC)(AB) > & , supp(S_2) = 0.10 \\ S_3 =< (AB)(CE) > & , supp(S_3) = 0.12 \\ S_4 =< (AB) > & , supp(S_4) = 0.13 \\ S_5 =< (EA)(BC) > & , supp(S_5) = 0.20 \end{cases}$$

Then,

$$LTC(S_1 =< (AB)(BC) >) = \frac{supp(S_1) - \sum\limits_{s_d \in S_d} supp(s_d)}{\sum\limits_{s_t \in S_t} supp(s_t)}$$
$$= \frac{0.25 - 0.10}{0.67} = 0.224$$

$$with \begin{cases} supp(S_1) = 0.25 \\ S_d = \{S_2\} \\ S_t = \{S_1, S_2, S_3, S_5\} \end{cases}$$

We find (BC) and (AB) in S_2 (which has the same itemsets as the sequence S_1, but in a different order) and found items A, B and C in S_2, S_3 and S_5, but not in S_4 which only contains items A and B.

Algorithm: Algorithm 1 describes the steps for computing LTC. This algorithm is divided into two steps: (1) we look for sequences containing common itemsets between the studied sequence and candidate sequence without considering the position of appearance (the first inner loop), and; (2) we look for all sequences from among candidate sequences that contain all items contained in the sequence under study (second inner loop). This algorithm has a complexity of at most $O(n^2 * m)$ where n is the number of sequential patterns in the sequential pattern database ($spDB$) obtained at the end of the mining step and m represents the size of the longest pattern.

3.5. Restitution of extracted patterns

In the extraction of knowledge, once the sequential patterns extraction process has been run, it is essential to compare the similarity of obtained objects (e.g., for sequential patterns visualization). This task is more difficult if we have results that include complex sequences composed of sets of items such as sequential patterns.

Algorithm 1: Calculation of the least temporal contradiction

Input: $spDB$: Database of sequential patterns and its supports

Output: *the least temporal contradiction* for each sequential pattern $S \in spDB$

```
begin
    LTC = φ;
    forall the (sequential pattern S1 ∈ spDB) do
        suppSd ← 0 ;
        suppSt ← 0 ;
        forall the (sequential pattern S2 ∈ spDB − {S1}) do
            all_in ← true ;
            while (all_in) do
                forall the (itemset IS ∈ S1) do
                    if (IS ⊈ S2) then
                        all_in ← false;
                    else
                        next IS ∈ S1;
                    end
                end
            end
            if (all_in) then
                suppSd ← suppSd + supp(S2) ;
            end
            all_in ← true ;
            while (all_in) do
                forall the (item I ∈ S1) do
                    if (I ∉ S2) then
                        all_in ← false;
                    else
                        next I ∈ S1;
                    end
                end
            end
            if (all_in) then
                suppSt ← suppSt + supp(S2);
            end
        end
        LTC(S1) ← (supp(S1) − suppSd) / suppSt ;
    end
    return LTC;
end
```

Several approaches have been developed to compare similarity between two sequences [37]. In this paper, we use a similarity measure called *Similarity Measure for Sequential Patterns* (S^2MP) proposed in [27] which takes into account the characteristics and semantics of sequential patterns. Indeed, this method can be used to measure the similarity of patterns and therefore provide coherent groups of analogous sequential patterns. S^2MP is based on two scores: the mapping score which takes into account the number of common itemsets between two sequences and the order score which takes into account the common order of these itemsets.

The mapping score S_m: We associate all itemsets of the first sequence with all itemsets of the second sequence and for each association we compute a weight (number of common items divided by the number of items of two compared itemsets divided by two). For each possible combination of associations, we calculate an average of weight and store the combination associated with the best average.

For example, consider two sequences $S_1 =< (A,B,C)(A,B)(C,D) >$ and $S_2 =< (A,B)(C,A)(A) >$. The weight associated with the association between the itemsets (A,B,C) in S_1 and (A,B) in S_2 is equal to $2/((3+2)/2) = 0.8$. Similarly, we associate (A,B,C) and (C,A) with a score of 0.8 and (A,B,C) and (A) with a score of 0.5. For the first itemset (A,B,C) of S_1, the association selected is (A,B) with a score of 0.8. The same procedure is adopted with the other itemsets in S_1: we combine the itemset (A,B) with (A) with a weight of 0.6 and the itemset (C,D) with the itemset (C,A) with a weight 0.5. Finally, S_m is the average weight of these three associations, i.e., 0.65.

The order score S_o: To calculate this score, we aggregate two subscores: *totalOrder*, the percentage of associations respecting sequence order and *positionOrder*, which correspond to the gap between two consecutive associations. To achieve this, we use the formula:

$$S_o = max\left\{totalOrder(sub) * positionOrder(sub)\right\} \quad (3)$$
$$with \text{ sub} \in \{ \text{ sub-sequences growing on studied sequence}\}.$$

For example : The order of itemsets associated to S_1 on S_2 are (1,3,2). We

find two growing sub-sequences, (1,3) and (1,2).

$$\begin{aligned} TotalOrder &= 2/3 \\ PositionOrder(\{1,3\}) &= 1-(1-2)/3 = 2/3 \\ PositionOrder(\{1,2\}) &= 1-(2-1)/3 = 2/3 \\ S_o &= max(\frac{2}{3}*\frac{2}{3};\frac{2}{3}*\frac{2}{3}) \\ &= 0.44 \end{aligned}$$

The value of S^2MP measure is the half of the product of mapping score and order score.

$$\begin{aligned} S^2MP &= S_m * S_o/2 \\ &= 0.65 * 0.44/2 \\ &= 0.143 \end{aligned}$$

This measure was used - simultaneously - to compare the patterns obtained by considering the three spatialization approaches.

4. Application to hydrological data

In this section, we describe the application of our spatiotemporal knowledge discovery process to hydrological data of the Saône watershed.

4.1. Context and data

Our database is composed of biological indicators measured on the Saône watershed. Figure 1 describes the geographical location of watercourses and water sampling stations in this watershed. Table 2 shows a small portion of the complete database.

Two types of data are available: (1) static informations related to the station itself (its location, its reference code, etc.) and; (2) dynamic informations which correspond to data measured by the station.

Static data concerns to the characteristics of each station, i.e.:

- Identification of the station (**codstace**)

Table 2: Data of Saône watershed

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

- Code of the surface water body where the station is located (**codmasseau**). Surface water body can correspond to a river, a canal, a section of a river or a section of a canal. For the Saône, there are 572 watercourses corresponding to surface water bodies. Objects like lakes and ponds are not studied;
- Spatial coordinates of the station (**x, y**). The Lambert Projection System 93 is used for the geo-referencing;
- Hydro-ecoregion code (**hydroecor**). A Hydro-ecoregion is a homogeneous spatial unit in terms of geology, topography and climate. This is one of the main criteria in the typology and definition of surface water bodies. Metropolitan France is divided into 22 hydro-ecoregions and 7 hydro-ecoregions are presents in the studied area;
- A kilometric point of the station on the watercourse. This measure, in kilometers, corresponds to the distance from the downstream confluence to the water quality station following watercourse;
- Size of water bodies at the station point. Sizes are ranking in five classes (very small, small,..., extra large) according to *Strahler* order that allows to define the spatial hierarchy of hydrographic network;
- Fish context of the station. This is a spatial unit for which a fish population operates independently.

Dynamic data are measures conducted by the stations. The frequency of these records varies with time and stations. Some stations have recurrent sample data while other stations only have a single sample data (e.g., for general monitoring). The main items associated with records are:

- Date of measure (**rdate**);
- Standardized Global Biological Index (**ibgn**). This index, called IBGN, is a standardized measure based on identification of macro-invertebrates in rivers;
- Biological Diatom Index (**ibd**). This index, named IBD, is a standardized measure to diagnostic trophic pollutions.

IBGN, IBD and a measure corresponding to the fusion of IBGN and IBD are standardized according to the water body and the hydro-ecoregion attributes. The IBGN and IBD measures have been made taking into account two points of view: a note (e.g., *idb_note*) and their current status (e.g., *ibgn_stat*). In addition, other three variables have been included in our dataset: (1) the taxonomic variety (*var_taxo*) representing the total number of taxa collected during a sampling, even if they are only represented by a single individual; (2) the faunal group that is the more sensitive to pollution (*gr_indic*), and; (3) the IBD measure established before the DCE regulation in France (*IBD2007*). All this information is used to estimate the condition of the watercourse at a specific survey point.

The data set consists of 12 features and $2,534$ rows. Table 3 describes the attributes A_i and their domain of values $dom(A_i)$.

Table 3: Description of attributes and their domain

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

4.2. Data pre-processing

In this section, we first present how data has been discretized. Then, we describe the spatial relationships used to materialize the spatial data in the

form of a non-spatial table. This transformation allows sequential pattern mining to be used to study hydrological evolutions at different level of spatial granularities, i.e. capturing the spatial information of data from two different points of view (c.f. Section 3.2).

4.2.1. Data discretization

Frequency histograms of each attribute are studied to discretize the data. In our case, the components are satisfactorily separated, and the number of observations sufficient. Thus, the frequency histogram provides a good estimation of the number of components and their values. Figures 6 and 7 show examples of frequency histograms for attribute *var_taxo* and *ibd* (respectively).

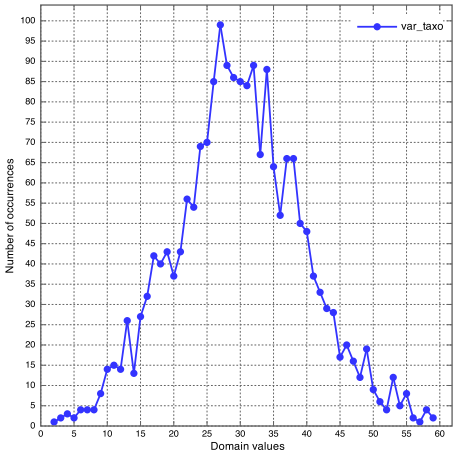


Figure 6: Frequency histogram for the *var_taxo* descriptor

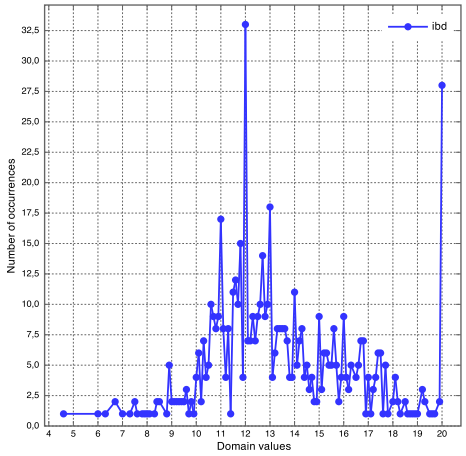


Figure 7: Frequency histogram for the *ibd* descriptor

Then, we apply a discretization based on an equal frequency. The values are partitioned such that each partition contains approximatively the same number of data. Using this type of discretization, we obtain balanced range of values. Final discretized values are presented in Table 4 with their corresponding attribute.

4.2.2. Mining to consider three distinct spatial relationships

Three hypotheses are studied to analyze the water quality status according to two bioindicators at different levels. Each hypothesis is closely related to the spatial relationship considered when pre-processing the data.

Table 4: Discretization of domain values $dom(A_i)$

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

1. The first hypothesis is the most naive one: "all stations are independent". Water quality is confined to the station area. It is not influenced by what is happening in the same watercourse, or in nearby ones. In this method, called **without zoning**, data is not pre-processed, and each record in the database corresponds to a zone. This approach constructs 711 zones (one for each monitoring stations).
2. The second hypothesis is: "stations of the same watercourse are linked". Water quality status in a particular station X may impact status of the other stations downstream in the same watercourse. Thus, we take account propagation and resilience processes. This division of space is called **watercourse** and generates 233 zones in our data set.
3. The third hypothesis is: "close stations are linked". Close stations to a given station X undergo the same kinds of pressures (e.g., in groundwater, in nearby agricultural areas,...). The area studied is a ϵ kilometers square around station X. This division of space, labeled **ϵ-neighborhood**, can be used to observe potentially indirect effects of pollution. 223 areas have been obtained using this approach.

Original data is mapped to sequences of values according to these spatial relationships. Data belonging to the same zone (w.r.t. chosen spatial relationship) are grouped into sets and ordered w.r.t. date, leading to sequences of itemsets. Thus, classical sequence mining algorithms can be used to extract *spatially frequent sequences* (sequential patterns). Table 5 shows the characteristics of three datasets.

Table 5: Characteristics of datasets

Datasets	Number of zones	Number of dates per zone (min/max)	Number of items per date (min/max)
without zoning	711	1/11	1/10
watercourse	233	2/37	1/10
ϵ-neighborhood	223	1/46	1/10

4.3. Mining sequential patterns

We use the PrefixSpan algorithm [33] for the extraction of sequential patterns because of its effectiveness at "mining" large volumes of data. This algorithm is based on the *pattern-growth* strategy used in [38]. The principle of this approach is to extract frequent patterns without a candidate generation step. This approach recursively creates a projected database, then associates it with a fragment of frequent pattern, and finally "mines" each projected database separately. Let S be a sequence of itemsets of the database DB. The database projection DB w.r.t. S, denoted $DB|_S$, is the set of suffixes of S in DB. With this approach, frequent patterns are extended progressively along a depth-first exploration of the search space.
For example, let the sequence database described in Table 6 and a minimum support $minsupp = 0.5$ for which we apply the PrefixSpan algorithm.

Table 6: Database of sequences

Id	Sequence
1	$< (A, B)(C) >$
2	$< (A, B)(A, B) >$
3	$< (A, E)(D, E) >$

Frequent 1-sequences are $< A >$ with a support of 1 and $< B >$ with a support of 0.66. $< C >$, $< D >$ and $< E >$ are not frequent. The database projections for these two sequences are presented in Table 7.

Finally, the frequent sequences occurring at least twice in the initial base sequences are: $< A >$, $< B >$ and $< (A, B) >$.

For our experiments, we have used *SPMF (Sequential Pattern Mining Framework)* [1]. We have extracted spatially frequent sequences in our dataset

[1] Available on `http://www.philippe-fournier-viger.com/spmf/`

Table 7: Projections of sequences table for the items $< A >$ and $< B >$

Prefix	Projected database	Frequent patterns
$< A >$	$< (\_, B)(C) >$	$< (A, B) >: 2$
	$< (\_, B)(A, B) >$	
	$< (\_, E)(D, E) >$	
$< B >$	$< (C) >$	
	$< (A, B) >$	

w.r.t. the three spatial relationships defined in Section 4.2.2:

1. *Without zoning*: The dataset consists of 711 sequences. Pattern mining was done with a minimum support threshold of 0.3. We have obtained 22 sequential patterns, all of size 1. Table 8 shows some solutions.

Table 8: Some patterns obtained with the *without zoning* approach.

Sequential patterns	Supp
<(ibgn_etat_TBE)>	0.32
<(ibgn_etat_TBE, ibgn_note_4)>	0.32
<(ibgn_0-10, gr_indic_0-4)>	0.32
<(ibgn_etat_BE, ibgn_note_3)>	0.31
...	...

2. *Watercourse zoning*: We applied the *PrefixSpan* algorithm to a dataset composed of 233 sequences with a minimum support of 0.3. We have obtained 564 sequential patterns, with 110 1-sequences (i.e., sequences of size 1), 361 2-sequences, 90 3-sequences and 3 4-sequences. Some of the solutions are presented in Table 9.

Table 9: Some patterns obtained with the *watercourse* approach.

Sequential patterns	Supp
<(ibgn_11-15) (ibgn_11-15, var_taxo_21-30)>	0.41
<(var_taxo_21-30) (var_taxo_21-30) (ibgn_11-15)>	0.36
<(ibgn_11-15, ibgn_etat_Emoy, ibgn_note_2) (ibgn_11-15)>	0.35
<(ibgn_note_2) (ibgn_etat_Emoy, ibgn_note_2)>	0.31
<(gr_indic_5-6, var_taxo_21-30, ibgn_etat_Emoy, ibgn_note_2)>	0.30
<(var_taxo_21-30) (ibgn_11-15, var_taxo_21-30) (var_taxo_21-30)>	0.33
<(var_taxo_21-30) (ibgn_16-20, var_taxo_31-40)>	0.30
...	...

3. *ϵ-neighborhood zoning*: We applied the same algorithm to a dataset consisting of 223 zones with a minimum support threshold of 0.3. We have obtained 138 1-sequences, 1,174 2-sequences, 658 3-sequences, 104 4-sequences and 8 5-sequences. In total, 2082 sequential patterns were extracted. Some of these patterns are presented in Table 10.

Table 10: Some patterns obtained with the *ϵ-neighborhood* approach.

Sequential patterns	**Supp**
<(var_taxo_21-30, ibgn_etat_Emoy)>	0.48
<(ibgn_16-20, gr_indic_7-9, var_taxo_31-40, ibgn_etat_TBE, ibgn_note_4)>	0.38
<(ibgn_note_2) (ibgn_etat_Emoy, ibgn_note_2)>	0.36
<(var_taxo_21-30, ibgn_note_2) (ibgn_11-15, var_taxo_21-30)>	0.35
<(gr_indic_7-9) (ibgn_16-20, gr_indic_7-9, ibgn_etat_TBE, ibgn_note_4)>	0.39
<(ibgn_etat_Emoy, ibgn_note_2) (var_taxo_21-30)>	0.42
<(var_taxo_21-30, ibgn_etat_Emoy) (var_taxo_21-30) (ibgn_11-15)>	0.31
<(var_taxo_21-30, ibgn_etat_Emoy, ibgn_note_2) (ibgn_11-15) (ibgn_11-15, var_taxo_21-30)>	0.31
<(var_taxo_21-30) (var_taxo_21-30) (ibgn_11-15)>	0.42
...	...

The support of the last sequence in Table 9 means that sequential pattern $<(var\_taxo\_21\text{-}30)(ibgn\_16\text{-}20, var\_taxo\_31\text{-}40)>$ appears for 30% of stations in the sequence database and can be interpreted as: taxonomic variety increases "very often" over time. The extracted patterns represent the evolution of a set of characteristics (biological indicators) belonging to a set of monitoring stations grouped using different spatial proximities.

The number of sequential patterns obtained using PrefixSpan algorithm on dataset with the three spatialization approaches and using a minimum support of 0.3 is respectively 22 without zoning, 564 with watercourse zoning and 2,082 with ϵ-neighborhood zoning. Interestingly, we obtained fewer patterns using the first approach than with the two other spatialization approaches.

4.4. Sequential patterns post-processing

This section describes the two post-processing steps performed on extracted sequential patterns. The first one filters sequential patterns according to the new *least temporal contradiction (LTC)* measure, thus leading to more pertinent spatiotemporal patterns. The second one groups relevant patterns into clusters using the S^2MP measure, leading to an easier interpretation of results by experts.

4.4.1. Application of the least temporal contradiction

We use the LTC measure to filter the most relevant sequential patterns. The LTC is computed as follows: let *spDB* be a database of sequential patterns obtained after running the PrefixSpan algorithm on the Saône watershed dataset - for example, using the watercourse approach - and given a sequence $S \in spDB$ and its support presented in Table 11.

Table 11: Sample sequence and its support

Sequential patterns	**Supp**
<(ibgn_16-20, ibgn_etat_TBE) (var_taxo_31-40)>	0.34

First, to calculate S_d, we look for sequences in $spDB$ containing itemsets $(ibgn\_16\text{-}20, ibgn\_etat\_TBE)$ and $(var\_taxo\_31\text{-}40)$ in a different position. We found two solutions (see Table 12), then S_d value for the sequence *<(ibgn_16-20, ibgn_etat_TBE) (var_taxo_31-40)>* is 0.66.

Table 12: Sequences used to calculate S_d

Sequential patterns	**Supp**
<(**ibgn_16-20, ibgn_etat_TBE**) (ibgn_11-15)(**var_taxo_31-40**)>	0.34
<(**var_taxo_31-40**) (**ibgn_16-20, ibgn_etat_TBE**)>	0.32

In a second time, S_t is calculated. Purposely, we look for items belonging to sequence $<(ibgn\_16\text{-}20, ibgn\_etat\_TBE)\ (var\_taxo\_31\text{-}40)>$ in all sequences in $spDB$ database. We found these items in sequences shown in Table 13, then the sum of S_t supports is equal to 3.34.

Table 13: Sequences used to calculate S_t

Sequential patterns	**Supp**
<(**ibgn_16-20, var_taxo_31-40, ibgn_etat_TBE**)>	0.36
<(**ibgn_16-20, var_taxo_31-40, ibgn_etat_TBE**, ibgn_note_4)>	0.36
<(**ibgn_16-20**, gr_indic_7-9, **var_taxo_31-40, ibgn_etat_TBE**)>	0.34
<(**ibgn_16-20**, gr_indic_7-9, **var_taxo_31-40, ibgn_etat_TBE**, ibgn_note_4)>	0.34
...	...

Finally, the least temporal contradiction (LTC) for the sequence $<(ibgn\_16\text{-}20, ibgn\_etat\_TBE)(var\_taxo\_31\text{-}40)>$ is:

$$LTC(<(ibgn\_16\text{-}20, ibgn\_etat\_TBE)(var\_taxo\_31\text{-}40)>) = \frac{0.34 - 0.66}{3.34}$$
$$= -0.09580838323353$$

The LTC measure was performed for the three sequential patterns datasets obtained w.r.t. pollution hypotheses. Tables 14, 15 and 16 show some sequences with their support (*Supp*) and least temporal contradiction (LTC), for the different spatialization approaches.

Table 14: LTC for data *without zoning*

Sequential patterns	**Supp**	**LTC**
<(ibgn_etat_TBE, ibgn_note_4)>	0.32	1.0
<(ibgn_11-15, var_taxo_21-30)>	0.39	1.0
<(var_taxo_21-30)>	0.5	0.1236
<(ibgn_0-10)>	0.36	0.05882
...	...	...

Table 15: LTC for data using the *watercourse* approach

Sequential patterns	**Supp**	**LTC**
<(var_taxo_21-30) (ibgn_16-20, var_taxo_31-40)>	0.3	1.0
<(ibgn_0-10, gr_indic_0-4, ibgn_etat_Emedio, ibgn_note_1)>	0.32	1.0
<(ibgn_0-10, ibgn_etat_Emedio, ibgn_note_1)>	0.34	0.0303
<(ibgn_note_1)	0.35	-0.738806
<(ibgn_note_4) (ibgn_etat_TBE)>	0.34	-0.963176
...	...	...

Table 16: LTC for data using the ϵ-*neighborhood* approach

Sequential patterns	**Supp**	**LTC**
<(ibgn_11-15) (ibgn_16-20, gr_indic_7-9)>	0.33	1.0
<(var_taxo_21-30) (ibgn_etat_TBE, ibgn_note_4)>	0.33	0.03125
<(gr_indic_7-9) (ibgn_11-15, var_taxo_21-30)>	0.36	0.01887
<(gr_indic_7-9) (ibgn_etat_BE, ibgn_note_3)>	0.31	-0.030928
<(var_taxo_21-30) (var_taxo_31-40)>	0.42	-0.215329
<(ibgn_etat_TBE, ibgn_note_4) (var_taxo_31-40, ibgn_note_4)>	0.31	-0.905918
...	...	...

In Table 16, we show some patterns found, their support and their least temporal contradiction measure using ϵ-neighborhood spatialization approach. In this table, sequential pattern $<(ibgn\_11\text{-}15)$ $(ibgn\_16\text{-}20, gr\_indic\_7\text{-}9)>$ appears for 33% of stations in the sequences database and it is never contradicted by the data over time. In contrast, sequential pattern $<(ibgn\_etat\_TBE, ibgn\_note\_4)$ $(var\_taxo\_31\text{-}40, ibgn\_note\_4)>$ appears for 31% of stations but it is very contradictory over time. Then, the first sequential pattern is more relevant.

To conclude, the *support* threshold allows us to extract the most spatially frequent patterns, additionally, the LTC measure enables the ranking of the most relevant spatiotemporal frequent patterns, i.e., those which are the least contradicted over time. The LTC measure can be directly exploited by the experts, for example bringing up patterns with a positive (w.r.t. negative) LTC value or showing the *top-k* patterns (for example the first 20 least contradicted sequential patterns).

4.4.2. Clustering of sequential patterns using S^2MP measure

To compare the similarity between the sequential patterns obtained, we have used the S^2MP measure to identify the irregularities and to construct homogeneous object classes. We have applied the S^2MP algorithm on two sets of results: (1) sequential patterns extracted using the watercourse approach, and; (2) sequential patterns obtained using the ϵ-neighborhood approach.

Table 17 shows some S^2MP distances for sequential patterns discovered using watercourse approach. It is important to notice that this measure evaluate sequences at two levels: (1) S^2MP compares itemsets and their position in the sequence and; (2) S^2MP compares the similarity of items in the itemsets.

Table 17: S^2MP distance for sequential patterns extracted using *watercourse* approach

Sequential patterns 1	**Sequential patterns 2**	**distance**
<(var_taxo_21-30) >	<(var_taxo_21-30, ibgn_etat_Emoy)>	0.9
<(var_taxo_21-30) >	<(gr_indic_5-6, var_taxo_21-30, ibgn_etat_Emoy)>	0.833333
<(var_taxo_21-30) >	<(ibgn_11-15, var_taxo_21-30, ibgn_etat_Emoy) >	0.833333
<(ibgn_11-15, gr_indic_7-9)>	<(ibgn_0-10, gr_indic_0-4) >	0.75
<(ibgn_11-15, gr_indic_7-9)>	<(var_taxo_31-40)(ibgn_11-15, var_taxo_21-30) >	1.0
<(ibgn_11-15, var_taxo_21-30, ibgn_note_3) >	<(ibgn_note_4) >	0.833333
<(gr_indic_7-9)(ibgn_16-20, var_taxo_31-40)>	<(gr_indic_7-9) >	0.75
...	...	...

A distance measure is widely used within another technique as a clustering. In this work, we have grouped sequential patterns through a certain number of clusters fixed *a priori*. For this, we will use a simplest unsupervised learning algorithm called k-medoids (based on k-means method). This technique aims to partition a number of patterns into k-medoid-based clusters in which each pattern belongs to the cluster with the nearest mean. In [27] the authors have adapted *k-means* algorithm to be used on sequential patterns. The selection of k value (number of clusters to be formed) for unsupervised methods based on k-means technic is a frequent problem in data clustering [39]. In contrast to BDScan, DENCLUE and OPTICS are examples of density based clustering algorithms where the value of k is automatically estimated. In k-means technique the correct choice of k is often ambiguous.

As an example, we consider sequential patterns mined using watercourse approach (overall 564 patterns). Sequential patterns shown in Table 18 have been identified as *medoids* of clusters (w.r.t. centroids in k-means) for k equal to 10. Many values of k have been tested in order to determine the k value. Nonetheless, we have fixed the k value to 10 for two main reasons: the *rule of thumb* value computed on the smaller number of obtained sequences is slightly higher than 10 and finally, visualization of clusters is more difficult to interpret using number k. It is important to notice that a *medoid* is a sequential pattern that appears frequently in the cluster under exploration, hence, these sequences may be regarded as interesting for the expert.

Table 18: Medoids for sequential patterns extracted using *watercourse* approach

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

For each of these sequential patterns - medoid of clusters -, we have S^2MP distances to other sequences positioned around. For instance, Ta-

ble 19 shows the S^2MP distance between medoid <(*var_taxo*_21-30)> and others sequences.

Table 19: Sequential patterns and their distance from medoid <(*var_taxo*_21-30)>

Sequential patterns	**Distance**
<(var_taxo_21-30, ibgn_etat_Emoy)>	0.9
<(gr_indic_5-6)>	0.75
<(gr_indic_5-6, var_taxo_21-30)>	0.9
<(gr_indic_5-6, var_taxo_21-30, ibgn_etat_Emoy)>	0.833333
<(ibgn_11-15, var_taxo_21-30)>	0.9
...	...

As we can see in Table 19, numerous sequential patterns have the same distance from the medoid. To reduce the number of patterns shown to experts, we have grouped sequential patterns having the same distance and represent them as a single entity. For instance, in Table 19, sequential patterns <(*var_taxo*_21-30, *ibgn_etat_Emoy*)>, <(*gr_indic*_5-6, *var_taxo*_21-30)> and <(*ibgn*_11-15, *var_taxo*_21-30)>, have the same distance - 0.9 - from medoid <(*var_taxo*_21-30)>, thus, they will be jointly represented. Table 20 shows the number of sequential patterns *(Coincidences)* having the same distance *(Distance)* from <(*var_taxo*_21-30)>.

Table 20: Distance and number of coincidences for cluster represented by <(*var_taxo*_21-30)>

Medoid	**Distance**	**Coincidences**
<(var_taxo_21-30)>	0.0	4
<(var_taxo_21-30)>	0.75	1
<(var_taxo_21-30)>	0.833333	16
<(var_taxo_21-30)>	0.9	5
<(var_taxo_21-30)>	1.0	2

Later, all sequential patterns included in cluster centered on sequential pattern <(*var_taxo*_21-30)> are displayed graphically considering the S^2MP measure. Sequence <(*var_taxo*_21-30)> is located in the center of plane and other sequential patterns, grouped by distance, are displayed around it.

Figure 8, represents sequential patterns for watercourse zoning where each cluster and their center is represented by a different colored dot. For instance,

sequential pattern <(*var_taxo*_21-30)> and sequential patterns belonging to the same cluster - grouped by distance - are represented by •.

In the same way, we have applied k-medoids algorithm for sequential patterns mined using ϵ-neighborhood approach (overall 2082 patterns). Sequential patterns shown in Table 21 have been identified as medoid of clusters using k-medoids algorithm for k equal to 10.

Table 21: Medoids for sequential patterns extracted using *ϵ-neighborhood* approach

Sequential patterns
<(gr_indic_5-6, ibgn_etat_Emoy, ibgn_note_2)>
<(ibgn_11-15,var_taxo_21-30, ibgn_etat_Emoy)>
<(ibgn_16-20, gr_indic_7-9, ibgn_etat_TBE, ibgn_note_4)(ibgn_16-20, var_taxo_31-40)>
<(ibgn_etat_Emoy, ibgn_note_2)(ibgn_etat_BE, ibgn_note_3)>
<(gr_indic_0-4)(var_taxo_31-40)>
<(ibgn_etat_BE, ibgn_note_3)(ibgn_16-20, ibgn_etat_TBE, ibgn_note_4)>
<(ibgn_11-15, gr_indic_7-9)(ibgn_16-20)>
<(ibgn_11-15, gr_indic_7-9)(var_taxo_21-30)>
<(ibgn_11-15, var_taxo_21-30)(ibgn_note_3)>
<(ibgn_11-15, gr_indic_5-6)(gr_indic_5-6)(ibgn_11-15, gr_indic_5-6)>

As we have done previously, to reduce the number of patterns shown to experts, we have grouped sequential patterns having the same distance and represent them as a single entity. For instance, sequence <(*gr_indic*_5-6, *ibgn_etat_Emoy*, *ibgn_note*_2)> has 52 sequences with a similarity distance of 0.75 (c.f. Table 22).

Table 22: Distance and number of coincidences for cluster represented by <(*gr_indic*_5-6, *ibgn_etat_Emoy*, *ibgn_note*_2)>

Center of cluster	**Distance**	**Coincidences**
<(gr_indic_5-6,ibgn_etat_Emoy,ibgn_note_2)>	0.0	5
<(gr_indic_5-6,ibgn_etat_Emoy,ibgn_note_2)>	0.75	52
<(gr_indic_5-6,ibgn_etat_Emoy,ibgn_note_2)>	0.833333	168
<(gr_indic_5-6,ibgn_etat_Emoy,ibgn_note_2)>	0.9	1

Figure 9 shows sequential patterns for *ϵ-neighborhood* zoning approach grouped by distance. Each cluster and their medoid have been represented by a different colored dot. For instance, medoid <(*gr_indic*_5-6, *ibgn_etat_Emoy*, *ibgn_note*_2)> represented by •, has sequential patterns strongly associated with it. In contrast, distances between

medoid $<(ibgn\_11\text{-}15, gr\_indic\_7\text{-}9)(ibgn\_16\text{-}20)>$ and sequences around it, - represented by ● - are near to 1. An early interpretation is: "the cluster of sequential patterns represented by ● is *less* interesting for the expert".

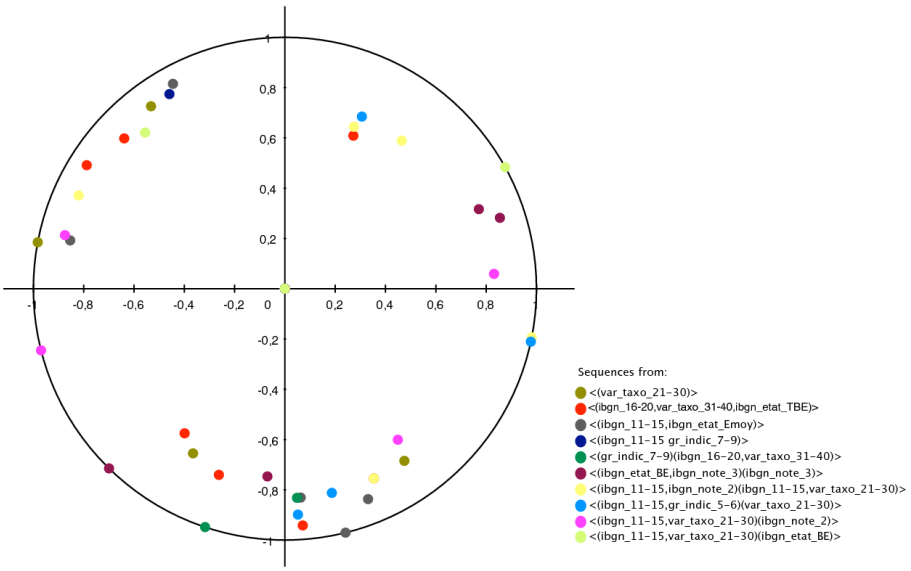


Figure 8: Distance between medoid and other sequential patterns - grouping by distance - for *watercourse* approach

Finally, Figure 10 shows the sequential patterns grouped by distance for both approaches, i.e., *watercourse* and ϵ-*neighborhood*. In this figure, we can note that the sequential patterns obtained using the watercourse approach are slightly closer to center of plane. This difference can be seen through the objective function called *sum of squared errors* or *SSE*. The SSE value for a given cluster is computed by: for each instance in the cluster, summing the squared differences between each attribute value and the corresponding one in the cluster medoid. These values are summed up for each instance in the cluster and for all clusters. The SSE value, which represent the cohesion of clusters, is equal to 22.4136 using watercourse approach which is smaller than the SEE value obtained for clusters using ϵ-neighborhood approach (around 25.3068). Indeed, the cohesion between instances using watercourse approach is strongest than other approach and consequently they are more interesting for experts.

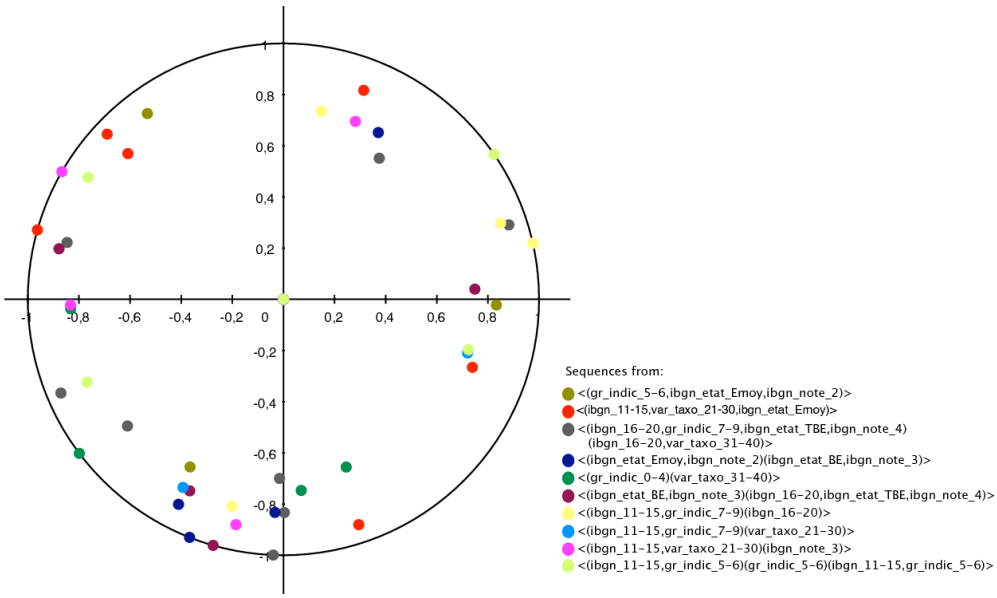


Figure 9: Distance between center and other sequential patterns - grouping by distance - for ϵ-*neighborhood* zoning approach

5. Conclusion and future research directions

In this paper, we have presented a knowledge discovery process on hydrological data. In particular, we have applied a conventional algorithm for sequential pattern extraction according to three spatialization approaches. We highlighted the problems that are posed regarding choices made in terms of spatialization and their influence on the number of extracted patterns. We have proposed an objective measure of validation: the *least temporal contradiction* measure which provides experts with an appropriate measure for the evaluation of obtained patterns. We also applied a similarity measure to compare sequences of patterns extracted using the different spatial proximities. Based on this measure, we display to experts coherent groups of analogous patterns.

This work has been conducted *blind*, i.e., without the intervention of data specialists. The results underline the difficulties involved in pre-processing search data without a thorough knowledge of the study area in question.

The perspectives for this work are numerous. First, regarding the data processed, additional elements to the determination of the pressures on water

are currently in the acquisition phase. Indeed, the exact determination of watercourse conditions requires other indicators that are absent from data presently studied. Therefore, other bioindicators or physic-chemical parameters are currently being acquired. Regarding the extraction phase, we would like to compare different data mining techniques in terms of obtained patterns.

Later, we will extend this approach by using pressure data, characterized by land use and survey data. The methodological issues are numerous: How to describe the pressures on watercourses based on land use data? How to model the relationship between land uses and river quality? And how to take into account data heterogeneity?

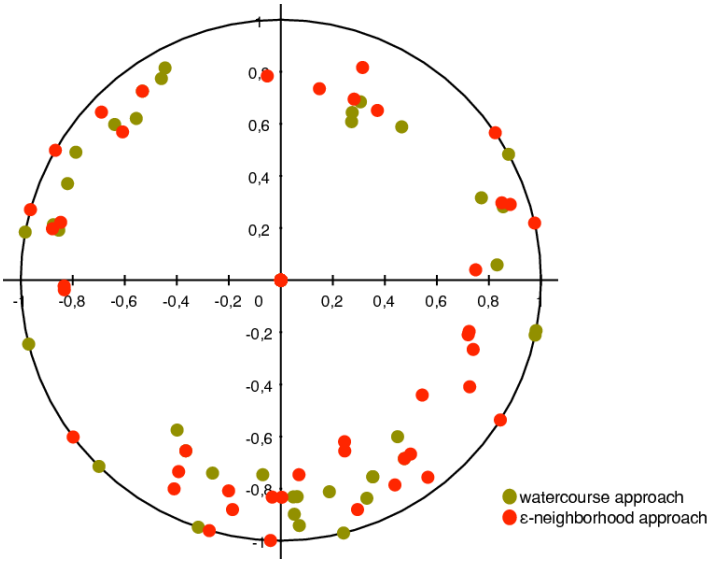


Figure 10: Deployment of sequential patterns - grouped by distance - around to medoids for *watercourse* and the ϵ-*neighborhood* approaches

References

[1] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, Advances in knowledge discovery and data mining, American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996, pp. 1–34.
URL `http://dl.acm.org/citation.cfm?id=257938.257942`

[2] O. Brazhnik, Databases and the geometry of knowledge, Data & Knowledge Engineering 61 (2), 2007, pp. 207–227.
URL http://www.sciencedirect.com/science/article/pii/S0169023X06000917

[3] L. Cao, H. Zhang, Y. Zhao, D. Luo, C. Zhang, Combined mining: Discovering informative knowledge in complex data, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 41 (3), 2011, pp. 699 –712. doi:10.1109/TSMCB.2010.2086060

[4] W. Hsu, M. Lee, J. Wang, Temporal and Spatio-Temporal Data Mining, Gale Virtual Reference Library, IGI Pub., 2008.
URL http://books.google.com/books?id=dpNyKGjM65cC

[5] G. Karina, I. Joaquín, H. Geoff, A. Ioannis, C. Joaquim, S.-M. Miquel, On the role of pre and post-processing in environmental data mining, in: The iEMSs: International Congress on Environmental Modeling and Software Integrating Sciences and Information Technology for Environmental Assessment and Decision Making, Vol. 3, 2008, pp. 1937–1958.

[6] V. Bogorny, P. Engel, L. O. Alvares, Spatial data preparation for knowledge discovery, IEEE Computer Graphics 24 (5), 2005, pp. 8.

[7] K. Koperski, J. Han, Discovery of spatial association rules in geographic information databases, in: Advances in Spatial Databases, Vol. 951 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 1995, pp. 47–66, 10.1007/3-540-60159-7_4.
URL http://dx.doi.org/10.1007/3-540-60159-7_4

[8] I. Tsoukatos, D. Gunopulos, Efficient mining of spatiotemporal patterns, in: Advances in Spatial and Temporal Databases, Vol. 2121 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2001, pp. 425–442, 10.1007/3-540-47724-1_22.
URL http://dx.doi.org/10.1007/3-540-47724-1_22

[9] B. Elias, Extracting landmarks with data mining methods, in: W. Kuhn, M. Worboys, S. Timpf (Eds.), Spatial Information Theory. Foundations of Geographic Information Science, Vol. 2825 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2003, pp. 375–389.

[10] M. Ester, H. P. Kriegel, J. Sander, Algorithms and applications for spatial data mining, Geographic Data Mining and Knowledge Discovery 5 (6), 2001.

[11] J. Mennis, J. W. Liu, Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change, Transactions in GIS 9 (1), 2005, pp. 5–17.

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

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

[14] J. Azé, Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances, in: Revue RIA-ECA numéro spécial EGC03, Vol. 17, 2003, pp. 171–182.

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

[16] L. Fleury, C. Djeraba, H. Briand, J. Philippe, Some aspects of rule discovery in data bases, in: S. Bhalla (Ed.), Information Systems and Data Management, Vol. 1006 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 1995, pp. 192–205. doi:10.1007/3-540-60584-3_32

[17] M. Jalali-Heravi, O. R. Zaïane, A study on interestingness measures for associative classifiers, in: Proceedings of the 2010 ACM Symposium on Applied Computing, SAC '10, ACM, New York, NY, USA, 2010, pp. 1039–1046. doi:http://doi.acm.org/10.1145/1774088.1774306

[18] Y. Huang, S. Shekhar, H. Xiong, Discovering colocation patterns from spatial data sets: a general approach, Knowledge and Data Engineering, IEEE Transactions on 16 (12), 2004, pp. 1472 – 1485. doi:10.1109/TKDE.2004.90

[19] P. Mohan, S. Shekhar, J. A. Shine, J. P. Rogers, Cascading spatio-temporal pattern discovery: A summary of results, in: Proceedings of the SIAM International Conference on Data Mining, SIAM, 2010, pp. 327–338

[20] Y. Huang, H. Xiong, S. Shekhar, J. Pei, Mining confident co-location rules without a support threshold, in: Proceedings of the 2003 ACM symposium on Applied computing, SAC '03, ACM, New York, NY, USA, 2003, pp. 497–501. doi:10.1145/952532.952630

[21] H. Mannila, P. Ronkainen, Similarity of event sequences, in: Temporal Representation and Reasoning, 1997. (TIME '97), Proceedings., Fourth International Workshop on, 1997, pp. 136 –139. doi:10.1109/TIME.1997.600793

[22] M. Capelle, C. Masson, J. franois Boulicaut, Mining frequent sequential patterns under regular expressions: a highly adaptive strategy for pushing constraints, in: In Proceedings SIAM DM 2003, Springer-Verlag, 2003, pp. 316–320.

[23] K. Sequeira, M. Zaki, Admit: anomaly-based data mining for intrusions, in: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '02, ACM, New York, NY, USA, 2002, pp. 386–395. doi:10.1145/775047.775103

[24] M. Vlachos, D. Gunopulos, G. Kollios, Robust similarity measures for mobile object trajectories, in: Proc. of DEXA Workshops, IEEE Computer Society, 2002, pp. 721–728.

[25] Q. Zhu, X. Wang, E. Keogh, S.-H. Lee, An efficient and effective similarity measure to enable data mining of petroglyphs, Data Mining Knowledge Discovery 23 (1), 2011, pp. 91–127. doi:10.1007/s10618-010-0200-z

[26] D. Bollegala, Y. Matsuo, M. Ishizuka, A web search engine-based approach to measure semantic similarity between words, Knowledge and Data Engineering, IEEE Transactions on 23 (7), 2011, pp. 977 –990. doi:10.1109/TKDE.2010.172

[27] H. Saneifar, S. Bringay, A. Laurent, M. Teisseire, S2MP: Similarity measure for sequential patterns, in: Proceedings of the 7th Australasian Data Mining Conference - Volume 87, 2008, pp. 95–104.

[28] E. Bertini, D. Lalanne, Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery, SIGKDD Explor. Newsl. 11 (2), 2010, pp. 9–18. doi:10.1145/1809400.1809404

[29] K. Wang, Y. Xu, J. X. Yu, Scalable sequential pattern mining for biological sequences, in: CIKM '04: Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, 2004, pp. 178–187. doi:http://doi.acm.org/10.1145/1031171.1031209

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

[31] C. Fiot, F. Masseglia, A. Laurent, M. Teisseire, Evolution patterns and gradual trends, International Journal of Intelligent Systems 24 (10), 2009, pp. 1013–1038. doi:10.1002/int.20369

[32] D.-A. Chiang, S.-L. Lee, C.-C. Chen, M.-H. Wang, Mining interval sequential patterns, International Journal of Intelligent Systems 20 (3), 2005, pp. 359–373. doi:10.1002/int.20070

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

[34] H. Cao, N. Mamoulis, D. W. Cheung, Mining frequent spatio-temporal sequential patterns., in: ICDM'05, 2005, pp. 82–89.

[35] M. Celik, S. Shekhar, J. P. Rogers, J. A. Shine, J. S. Yoo, Mixed-drove spatio-temporal co-occurence pattern mining: A summary of results., in: ICDM'06, 2006, pp. 119–128.

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

[37] T. Bie, Subjective interestingness in exploratory data mining, in: A. Tucker, F. Hppner, A. Siebes, S. Swift (Eds.), Advances in Intelligent Data Analysis XII, Vol. 8207 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2013, pp. 19–31. doi:10.1007/978-3-642-41398-8_3

[38] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, M.-C. Hsu, Freespan: frequent pattern-projected sequential pattern mining, in: Proc. of ACM SIGKDD, KDD '00, ACM, New York, NY, USA, 2000, pp. 355–359. doi:http://doi.acm.org/10.1145/347090.347167

[39] Y. Zhang, E. Cheng, An optimized method for selection of the initial centers of k-means clustering, in: Z. Qin, V.-N. Huynh (Eds.), Integrated Uncertainty in Knowledge Modelling and Decision Making, Vol. 8032 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2013, pp. 149–156. doi:10.1007/978-3-642-39515-4_13