Finding Relevant Sequences With The Least Temporal Contradiction Measure: Application to Hydrological Data

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

Multidisciplinary Research on Geographical Information in Europe and Beyond
Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, April, 24-27, 2012
ISBN: 978-90-816960-0-5
Editors: Jérôme Gensel, Didier Josselin and Danny Vandenbroucke

Finding Relevant Sequences With The Least Temporal Contradiction Measure: Application to Hydrological Data

Hugo Alatrista Salas
Irstea-TETIS, PPME
500, rue J. F. Breton, 34093
Montpellier, France
hugo.alatrista-salas@teledetection.fr

Jérôme Azé
LRI, Equipe Bioinformatique
Université Paris-Sud, 91405 Orsay Cedex
Paris, France
jerome.aze@lri.fr

Sandra Bringay
LIRMM, 161, rue Ada, 34392
Département MIAp-UM3, Route de Mende
34000, Montpellier, France
sandra.bringay@lirmm.fr

Frédéric Flouvat, Nazha Selmaoui-Folcher
Université de la Nouvelle Calédonie
PPME - BP R4, 98851
Nouméa, Nouvelle Calédonie
firstname.lastname@univ-nc.nc

Flavie Cernesson, Maguelonne Teisseire
Irstea-TETIS
500, rue J. F. Breton, 34093
Montpellier, France
firstname.lastname@teledetection.fr

Abstract

In this paper, we present a knowledge discovery process applied to hydrological data. To achieve this objective, we apply an algorithm to extract sequential patterns on data collected at stations located along several rivers. The data is pre-processed in order to obtain different spatial proximities and the number of patterns is estimated to highlight the influence of defined spatial relationship. We provide an objective measure of assessment, called *the least temporal contradiction*, to help the expert in discovering new knowledge. Such elements can be used to assess spatialized indicators to assist the interpretation of ecological and rivers monitoring pressure data.
Keywords: Data Mining, Patterns, Rivers, Spatial Mining.

1 Introduction

The water system, structuring landscapes and ecosystems of metropolitan France, covers more than 500000 km. The river system is a fragile environment subject to the presence of many economic activities and usages that have changed over time, and which have altered the physico-chemical and biological quality of water.

However, new European and French regulations demand the preservation and restoration of rivers and their surrounding environments. If systems for monitoring water quality have been existing for several decades, the challenge is now to construct indicators to take into account the influence of uses and restoration measures on the water quality.

To build an efficient tool, we must consider different types of data: (1) hydrological data, here, related to water quality (2) data related to monitoring stations (location, specific network...) (3) hydrographic network data, its physical characteristics and spaces associated with it: watershed, water mass ... (4) data related to human activities, and finally (5) pressure variables or context variables such as weather data, or data related to hydroecological homogeneity (such as hydro-ecoregions).

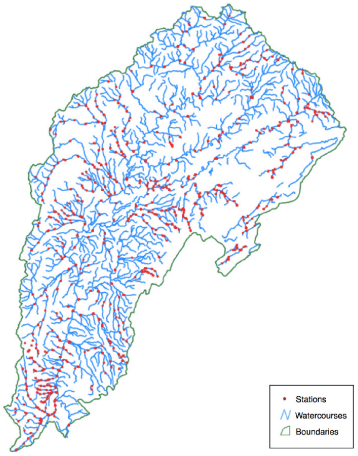
The data constitutes an important pool of information that is difficult to globally analyze. The heterogeneity and quantity of data handled requires the definition of specific approaches. This requirement is particularly important as we must take into account the temporal variability of data.

In this paper, we aim at analyzing the water quality in the hydrological network of Saône (Burgundy) watershed. Our contribution should take into account the temporal variability of the data while considering the spatial proximity of different stations by grouping them according to their distance, to their membership in a common area, etc. For this, we address the overall process of knowledge discovery on hydrological data.

The data is pre-processed w.r.t. different spatial proximities. Then, we use a pattern mining algorithm to extract frequent temporal sequences. Finally, to help the expert in discovering new knowledge, we evaluate extracted sequences against a new measure, called *the least temporal contradiction*. This measure enables to find relevant sequences that are the least contradicted over time.

In section 2, we review the concepts of sequential pattern mining and define the Least Temporal Contradiction measure. Then, in section 3, we describe the knowledge discovery process applied on Saône watershed data and the spatial features that have been considered. Findings with the least temporal contradiction measure on extracted patterns are presented. The paper ends with our conclusions and some perspectives.

Figure 1: The Saône river watershed.



2 Definitions and methods

In this section, we give some preliminary definitions on sequential patterns and define the least temporal contradiction measure.

2.1 Preliminaries

Consider the database *DB*, illustrated in Table 1, which groups all records made by stations dispersed along several rivers (e.g. in Table 1, item *A* could be "good biological indicator IBGN").

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

Let $I = \{i_1, i_2, \ldots, i_m\}$ the set of *items* (quality status). 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 an non-empty ordered list, of itemsets denoted by $< s_1, s_2, \ldots, s_p >$ where s_j is an *itemset*.

A *n-sequence* is a sequence of n itemsets. For example, consider quality status 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 the Table 1. This means quality status 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 $< s_1, s_2, \ldots, s_p >$ is a subsequence of another sequence $< s'_1, s'_2, \ldots, s'_m >$ if there exist integers $k_1 < \ldots < k_j < \ldots < k_p$ such as $s_1 \subseteq s'_{k_1}, s_2 \subseteq s'_{k_2}, \ldots, s_p \subseteq s'_{k_p}$. For example, the sequence $S' =< (B)(E) >$ 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 itemsets of S. All quality status 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 his 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 that satisfies the minimum support (i.e. whose support is greater than *minsupp*) is a *frequent sequence* called a *sequential pattern*.

Table 1: Example of river quality status dataset

Client	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)

2.2 Sequential Patterns Mining

The problem of mining sequential patterns was introduced by [1] in the context of the basket market problem and applied with success in many fields such as biology [14, 13], Web mining [10, 8], anomaly detection [12], the data flow mining [7] or the description of behavior into group [11].

To extract sequential patterns, the *PrefixSpan* algorithm [9] has been adopted because of its effectiveness in large volumes of data. This method uses a *divide and conquer* strategy by performing a *depth-first search* with successive database projections.

A projection of the database according to a sequence S is defined by the set of suffixes of sequences present in the database and prefixed by S. The goal is to reduce the search space. In this context, *PrefixSpan* analyzes shared prefixes which are present in the data sequences. From this analysis, the algorithm builds intermediate databases (from the original database) that are projections deducted from identified prefixes. Then, for each intermediary database, *PrefixSpan* seeks to grow the set of sequential patterns discovered by applying the same process recursively.

2.3 The least temporal contradiction

In the data mining domain, it is common to obtain frequent sequences that are more numerous than in the original data. Choosing the most relevant sequences remains problematic since it is often closely linked to the data handled. Even if spatiotemporal data mining received a lot of attention [4, 5, 6], to our knowledge, there is no work on filtering the most relevant frequent sequences that are not contradicted over time.

For this, we propose to extend the measure called *the least contradiction (LTC)*, defined for association rules in [2] for two main reasons. First, this measure is simple to understand by experts and to implement. Second, previous work has exhibited the capacity of this measure to extract nuggets of knowledge [2] and to resist to noise [3]. Other measures could also be extended to temporal sequences such as *lift* if the definition is close to the one of the least contradiction.

Let S be a frequent sequence, the *Least Temporal Contradiction* of S, denoted $LTC(S)$, is defined by:

$$LTC(S) = \frac{supp(S) - \sum_{s_d \in S_d} supp(s_d)}{\sum_{s_t \in S_t} supp(s_t)}$$

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

The extension of the least temporal contradiction allows us to keep the original spirit of the measure which was designed to estimate how many times a rule is verified *vs* how many times it is disabled. A rule that is most frequently tested as disabled is a priori irrelevant. Like the *conventional* version, this measure 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 relevant sequence.

For instance, consider the following sequences and their support:

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

Then,

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

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

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_1, S_2, S_3 and S_5, but not in S_4 which only contains items A and B.

In the next section, we describe the knowledge discovery process used on hydrological data.

3 Knowledge Discovery for hydrological data

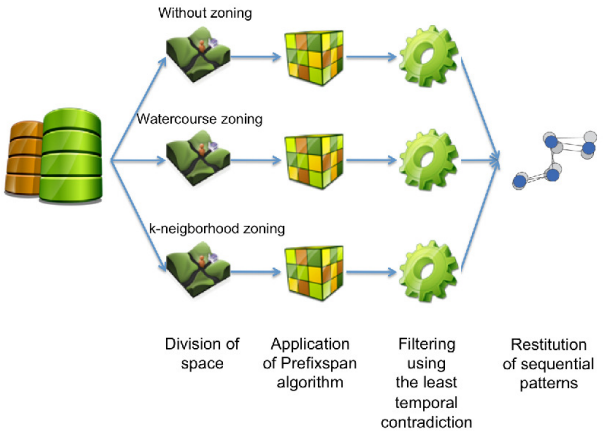
In this section, we describe the different steps, illustrated in Figure 2, that have been addressed in the process of knowledge discovery for the Saône watershed data. The pre-processing step consists in transforming data by grouping stations considering their different spatial proximity. In the pattern mining step, we use an algorithm to extract sequential patterns in order to take into account the temporal aspect. Finally, for the post-processing step, we define a new measure called *the least temporal contradiction* in order to find relevant sequences that are the least contradicted over time.

3.1 Data description

The database is composed of biological indicators measured on the Saône river and its tributaries. Figure 1 describes the geographical location of watercourses and weather stations in the considered watershed.

The data are static information related to the station itself (its location, its reference code, etc.) and dynamic information which corresponds to data recorded by the station equipment.
Static data is related to weather stations recorded on the waterways. Each station is described by:

Figure 2: Process of knowledge discovery applied to hydrologic network.



- Lambert coordinates (x, y): to identify the spatial position of each sampling station identified by *codstace*. The Lambert Projection System 93 is adopted here to perform the geo-referencing;
- A kilometric point: quantity used to locate a point along a watercourse which is calculated by measuring, in kilometers, the portion of the course between the located point and a point serving as origin (the confluence);
- A hydro-ecoregion: homogeneous spatial unit in terms of geology, topography and climate. This is one of the main criteria in the typology and definition of masses of surface water. Metropolitan France is divided into 22 hydro-ecoregions and 7 hydro-ecoregions are present on the study area;
- Codmasseau: to codify water masses, here corresponding to surface water such as rivers, canals, a section of a river or a section of a canal. For the Saône, there are 572 water courses type water masses. However, we do not treat water masses such as lakes and ponds;
- The size of water masses (Very Small, Small, ..., Extra Large) is based on physical dimension (hydraulic geometry, flow rates, watershed ground surface...).
- Fish context: spatial unit in which a fish population operates independently.

Dynamic data correspond to surveys 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 single sample data e.g. for ad-hoc studies. The main items associated with records are the following:

- The date of statement;
- The IBGN: Standardized Global Biological Index (standardized calculation based on identification of macroinvertebrates living in rivers);
- IBD: Biological Diatom Index (standardized calculation of diagnostic of trophic pollution).

Indicators IBGN and IBD are standardized according to the mass of water and the hydro-ecoregion studied.

Therefore, three notes are obtained and are comparable between the different stations: one note for IBGN, a note for IBD

and a note corresponding to the fusion of two normalized previous notes. This last information is used to estimate the condition of the watercourse at the point of survey.

3.2 Integration of spatial dimension

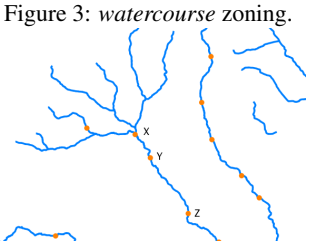
Spatial data can be used to determine the relevant geographic areas to manage (a) the proximity from the location of the stations expressed by their Lambert coordinates (geo-referenced coordinate system), (b) flow aspects combining the proximity related to water course, the flow direction and the connections between the rivers.

We pre-processed data to divide the space into zones. The data sequences are then obtained by combining data from the same area and sorting them by date. Thus, we can use a conventional algorithm to extract frequent sequences (sequential patterns), as explained in Section 2.2.

In this paper, two spatial division techniques have been used:

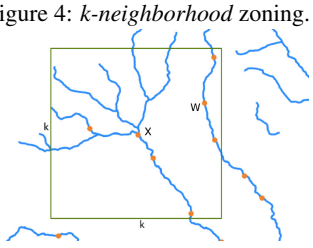
- A restricted neighborhood to *watercourse* : for a given watercourse, two stations X and Y located on this watercourse are considered as neighbors. For example in Figure 3, stations X, Y and Z belong to the same watercourse. These stations are considered as a single area, and their data are combined together.

Figure 3: *watercourse* zoning.



- The k-neighborhood: the space is divided into areas around each station by exploiting the Lambert coordinates. In each of these areas, stations that are located within an area of k km^2 centered on station X are grouped, even if these stations belonging to different watercourses. For example in Figure 4, stations X and W are considered to be in the same area, even if they are not on the same watercourse.

Figure 4: *k-neighborhood* zoning.



Thanks to these two spatial division methods, we are able to bring together the stations within areas and thus to aggregate data in order to extract frequent sequences. This provides the most relevant sequential patterns with regards to heterogeneous nature of the records.

These two approaches were used to obtained different hypotheses for the influence of pollution.

(i) The first hypothesis is that pollution measured in a given watercourse at a particular station X will have a potential impact on the downstream stations of X and on the other hand that the origin of the pollution is related to a phenomenon located upstream of X. The division by *watercourse* enables us to evaluate this hypothesis using *average* indicators of pollution for the whole river.

(ii) The second hypothesis is that the pollution measured at a station X is the result of pollution whose origin can be related to the same groundwater source, or in neighboring agricultural areas etc. The division into *k-neighborhood* is used to average the indicators of pollution in large areas in order to observe non local effects of watercourse associated with station X.

3.3 Mining sequential patterns

To perform experiments, we used *SPMF (Sequential Pattern Mining Framework)* implemented by Phillipe Fournier-Viguera and available from `http://www.philippe-fournier-viger.com/spmf/`.
We extract frequent sequential patterns in our data set using the following spatial division approaches:

1. *Without zoning*: The data set consists of 711 sampling stations identified by station identification code (codstace). The extraction was done with a minimum support of 0.3. We obtained 22 frequent patterns. The size of all these patterns is 1. Table 2 shows some the extracted patterns;

Table 2: Some patterns obtained with the *no zoning* approach.

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 neighborhood*: We applied the algorithm to a data set consisting of 233 zones with a minimum support of 0.3. We obtained 564 frequent patterns, with 110 1-sequences, 361 2-sequences, 90 3-sequences and 3 4-sequences. Some of the results found are presented in Table 3;

Table 3: Some patterns obtained with the *watercourse* approach.

Patterns	Supp
<(gr_indic_5-6, var_taxo_21-30, ibgn_etat_Emoy, ibgn_note_2)>	0.3
<(ibgn_note_2) (ibgn_etat_Emoy, ibgn_note_2)>	0.31
<(ibgn_11-15, ibgn_etat_Emoy, ibgn_note_2) (ibgn_11-15)>	0.35
<(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
<(var_taxo_21-30) (ibgn_11-15, var_taxo_21-30) (var_taxo_21-30)>	0.33
<(ibgn_11-15, var_taxo_21-30) (ibgn_11-15) (ibgn_11-15, var_taxo_21-30)>	0.31
...	...

3. *K-neighborhood*: We applied the same algorithm to a data set consisting of 223 zones with a minimum support of 0.3. We obtained 138 frequent patterns of size 1, 1174 frequent patterns of size 2, 658 of size 3, 104 of size 4 and 8 patterns of size 5. In total, 2082 frequent patterns were extracted. Some of these patterns are presented in Table 4.

Table 4: Some patterns obtained with the *k-neighborhood* approach.

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 number of sequences obtained by *PrefixSpan* algorithm on the data set with the three spatialization approaches, and with a minimum support set to 0.3 is respectively 22 without zoning, 564 with *watercourse* zoning and 2082 with *k-neighborhood* zoning.

It is interesting to highlight that we obtained few patterns using the first approach, unlike with the two other spatialization approaches.

3.4 Patterns and the least temporal contradiction measure

We have applied the *least temporal contradiction* measure to find the more relevant sequential patterns. Indeed, even if in terms of volume of processed data, a complete validation can be envisaged. This will not be the case when the data set is extended nationally.

The *least temporal contradiction LTC* was calculated as follows:

Let *SPDB* be a database of sequential patterns obtained after running the *PrefixSpan* algorithm [9] on the Saône river watershed data set by considering the spatialization of a *watercourse*, for example. Given a sequence $S \in SPDB$ and its support presented in Table 5:

Table 5: Sample sequence and its support

Sequence	Supp
$< (ibgn\_16\text{-}20, ibgn\_etat\_TBE)(var\_taxo\_31\text{-}40) >$	0.34

To calculate S_d, we looked for itemsets $(ibgn\_16\text{-}20, ibgn\_etat\_TBE)$ and $(var\_taxo\_31\text{-}40)$ in all sequences S of the database without considering the position which they appear in S. We find two solutions (see Table 6):

Table 6: Sequences used to calculate S_d

Sequence	Supp
$< (ibgn\_16\text{-}20, ibgn\_etat\_TBE)(ibgn\_11\text{-}15)(var\_taxo\_31\text{-}40) >$	0.34
$< (var\_taxo\_31\text{-}40)(ibgn\_16\text{-}20, ibgn\_etat\_TBE) >$	0.32

Finally, the value S_d for the sequence *<(ibgn_16-20, ibgn_etat_TBE) (var_taxo_31-40)>* is 0.66.

The calculation of S_t is performed in a similar way to the calculation of S_d. We seek the *items* belonging to the sequence $< (ibgn\_16\text{-}20, ibgn\_etat\_TBE)\ (var\_taxo\_31\text{-}40) >$ in all sequences S of sequential patterns database *SPDB*. We found these *items* in sequences shown in Table 7:

The sum of supports of the sequences $s_t \in S_t$ is equal to 3.34.

Table 7: Sequences used to calculate S_t

Sequence	Supp
$< (ibgn\_16\text{-}20, gr\_indic\_7\text{-}9, var\_taxo\_31\text{-}40, ibgn\_etat\_TBE) >$	0.34
$< (ibgn\_16\text{-}20, gr\_indic\_7\text{-}9, var\_taxo\_31\text{-}40, ibgn\_etat\_TBE, ibgn\_note\_4) >$	0.34
$< (ibgn\_16\text{-}20, var\_taxo\_31\text{-}40, ibgn\_etat\_TBE) >$	0.36
$< (ibgn\_16\text{-}20, var\_taxo\_31\text{-}40, ibgn\_etat\_TBE, ibgn\_note\_4) >$	0.36
…	…

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 objective measure of evaluation *LTC* was applied to the patterns obtained when running the selected algorithm on the Saône data with three proposed spatial approaches. Tables 8, 9 and 10 show some sequences and the values associated with the support value (*Supp*) and the value of the *least temporal contradiction (LTC)* measure for these different spatialization approaches.

Table 8: LTC for data *without zoning*

Sequence	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 9: LTC for data using the *watercourse* approach

Sequence	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_4) (ibgn_etat_TBE)>	0.34	-0.963176
<(ibgn_note_1)	0.35	-0.738806
…	…	…

Table 10: LTC for data using the *k-neighborhood* approach

Sequence	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
<(ibgn_etat_TBE, ibgn_note_4) (var_taxo_31-40, ibgn_note_4)>	0.31	-0.905918
<(var_taxo_21-30) (var_taxo_31-40)>	0.42	-0.215329
<(gr_indic_7-9) (ibgn_etat_BE, ibgn_note_3)>	0.31	-0.030928
…	…	…

To conclude, the *support* threshold allows to extract the most frequent patterns. The *LTC* measure enables to rank the most relevant frequent patterns, i.e. those which are the least contradicted.

4 Conclusion and perspectives

In this paper we have presented the first steps of a data mining project on hydrological data. In particular, we applied a conventional algorithm for sequential pattern extraction according to three spatialization approaches. We highlighted the problems

that are posed depending on 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 to experts an appropriate measure for the evaluation of obtained 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 of this work are numerous. First, regarding the data processed, additional elements on water pressures are currently in acquisition phase. Indeed, the exact determination of the condition of the watercourse requires other indicators that are absent from the data presently studied. Therefore the IPR (Fish River Index) and IBMR (Macrophytes River Biological Index) are currently being acquired. Then, for the extraction phase, we would like to compare different data mining techniques in terms of obtained patterns. Then, we will extend this approach by using pressure data, characterized by the 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 the heterogeneity of the data?

References

[1] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In Philip S. Yu and Arbee L. P. Chen, editors, *Proceedings of the Eleventh International Conference on Data Engineering, March 6-10, 1995, Taipei, Taiwan*, pages 3–14. IEEE Computer Society, 1995.

[2] Jérôme Azé. Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances. In *Revue RIA-ECA numéro spécial EGC03*, volume 17, pages 171–182, 2003.

[3] Jérôme Azé, Philippe Lenca, Stéphane Lallich, and Benoît-Vaillant. A study of the robustness of association rules. In R. Stahlbock, S. F. Crone, and CSREA Press S. Lessmann, editors, *The 2007 International Conference on Data Mining (DMIN'07)*, pages 132–137, 2007.

[4] Huiping Cao, Nikos Mamoulis, and David W. Cheung. Mining frequent spatio-temporal sequential patterns. In *ICDM'05*, pages 82–89, 2005.

[5] Mete Celik, Shashi Shekhar, James P. Rogers, James A. Shine, and Jin Soung Yoo. Mixed-drove spatio-temporal co-occurence pattern mining: A summary of results. In *ICDM'06*, pages 119–128, 2006.

[6] Loic Mabit, Nazha Selmaoui-Folcher, and Frédéric Flouvat. Modélisation de la dynamique de phénomènes spatio-temporels par des séquences. In *EGC*, pages 455–466, 2011.

[7] Alice Marascu and Florent Masseglia. Mining sequential patterns from data streams: a centroid approach. *Journal of Intelligent Information Systems*, 27(3):pages 291–307, 2006.

[8] Florent Masseglia, Pascal Poncelet, Maguelonne Teisseire, and Alice Marascu. Web usage mining: extracting unexpected periods from web logs. *Data Mining and Knowledge Discovery (DMKD)*, 16(1):pages 39–65, 2008.

[9] Jian Pei, Jiawei Han, B. Mortazavi-Asl, Jianyong Wang, H. Pinto, Qiming Chen, U. Dayal, and Mei-Chun Hsu. Mining sequential patterns by pattern-growth: the prefixspan approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):pages 1424–1440, nov. 2004.

[10] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, and Hua Zhu. Mining access patterns efficiently from web logs. In Takao Terano, Huan Liu, and Arbee L. P. Chen, editors, *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, pages 396–407. Springer, 2000.

[11] Dilhan Perera, Judy Kay, Irena Koprinska, Kalina Yacef, and Osmar R. Zaïane. Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, 21(6):pages 759–772, 2009.

[12] Julien Rabatel, Sandra Bringay, and Pascal Poncelet. Aide à la décision pour la maintenance ferroviaire préventive. In Sadok Ben Yahia and Jean-Marc Petit, editors, *Extraction et gestion des connaissances (EGC'2010), Actes, 26 au 29 janvier 2010, Hammamet, Tunisie*, Revue des Nouvelles Technologies de l'Information, pages 363–368. Cépaduès-Éditions, 2010.

[13] Paola Salle, Sandra Bringay, and Maguelonne Teisseire. Mining discriminant sequential patterns for aging brain. In Carlo Combi, Yuval Shahar, and Ameen Abu-Hanna, editors, *Artificial Intelligence in Medicine, 12th Conference on Artificial Intelligence in Medicine, AIME 2009, Verona, Italy, July 18-22, 2009. Proceedings*, Lecture Notes in Computer Science, pages 365–369, 2009.

[14] Ke Wang, Yabo Xu, and Jeffrey Xu Yu. Scalable sequential pattern mining for biological sequences. In *CIKM '04: Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management*, pages 178–187, New York, NY, USA, 2004. ACM.

Multidisciplinary Research on Geographical Information in Europe and Beyond
Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, April, 24-27, 2012
ISBN: 978-90-816960-0-5
Editors: Jérôme Gensel, Didier Josselin and Danny Vandenbroucke