A Spatial-based KDD Process to Better Understand the Spatiotemporal Phenomena

Hugo Alatrista Salas, Sandra Bringay, Frédéric Flouvat, Nazha Selmaoui-Folcher, Maguelonne Teisseire

A Spatial-based KDD Process to Better Understand the Spatiotemporal Phenomena

Hugo Alatrista-Salas

Irstea - TETIS 500, rue J. F. Breton, 34093 Montpellier, France
hugo.alatrista-salas@teledetection.fr
UNC - PPME BP R4, 98851 Noumea , New Caledonia
hugo.alatrista@univ-nc.nc

Abstract. In this paper, 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 two algorithms to extract spatiotemporal patterns and compare them. Such elements can be used to assess spatialized indicators to assist the interpretation of ecological and rivers monitoring pressure data.

Keywords: data mining, sequential patterns, spatiotemporal data

1 Introduction

In response to the rapidly rising and widespread use of database technology - including heterogeneous, geo-referenced and multidimensional data -, there is a growing interest in developing new techniques for extracting knowledge from data. These techniques are the subject of the emerging field of Knowledge Discovery in Databases (KDD). The KDD process is defined as the multi-steps process of discovering valid, novel, and potentially useful knowledge from large databases [7].

The KDD process can be very complex and the steps may change significantly depending on data origin. For instance, when data are geo-referenced, KDD process is considerably limited because spatial information is not easy to discern and can provide additional information. Currently, research in geographic knowledge discovery (i.e., KDD on spatial databases) is very active [17, 3]. However, the spatial characteristics of data are not fully exploited in the KDD process. For instance, river pollution study may lead to different space division following different pollution hypotheses. Knowing the impact of spatiality on handle strategies is essential to restore the ecological status of aquatic environments.

In this context, this paper focus on the impact of spatial components of data on the KDD process. For this, we propose a KDD process including two steps: first, we pre process the data in order to divide the space using two original spatialization approaches. Finally, we apply two different data mining techniques enabling to extract two semantically different kinds of spatiotemporal patterns. These techniques have been compared and differences were widely discussed in this paper.

This paper shows the current state of a PhD thesis, which focus on understand and modelize spatiotemporal phenomena. This thesis is under direction of Prof. Maguelonne Teisseire, Prof. Nazha Selmaoui-Folcher, Prof. Sandra Bringay and Prof. Frédéric Flouvat.

Problem statement

The water system, structuring landscapes and ecosystems of metropolitan France, covers more than 500000 km. divided into 6 water supply agencies containing several watersheds[1]. This structure is a fragile environment subject to the presence of many economic activities and usages that have increased the vulnerability of the water resources including rivers, canals, lakes, etc. In this context, river pollution is a phenomenon that is observed by measuring physicochemical and biological indicators for water quality. These indicators, which evolve over time and depend explicitly on the location of sampling stations strategically located along several rivers.

Two types of data are available: (1) static informations related to the station itself, e.g., its location (coordinates x, y), its reference code, etc., and; (2) dynamic informations which correspond to data measured by the station, e.g., the Standardized Global Biological Index (ibgn), Biological Diatom Index (ibd), the taxonomic variety (taxovar), the fish index (fishindex), etc.

This manuscript is organized as follows: In Section 2, we present our motivation and a detailed overview of the related work. Then, in Section 3, we describe our generic knowledge discovery process. Later, we apply our proposition on dataset and some patters obtained are shown. This paper ends with our conclusions and some perspectives.

2 Motivation

Knowledge discovery in databases (*KDD*) is a dynamic research field. In [7], authors presented the most widely used KDD framework and provided a broad overview of knowledge discovery techniques. Here KDD, was described as a set of interactive and iterative steps: data selection, pre-processing, transformation, data mining, and post processing or interpretation. As mentioned in [7], *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 one step of this general process. Indeed, if only data mining is used, this can lead to the discovery of meaningless patterns for experts.

In addition, the advent of GIS (Geographical Information Systems) technology and the availability of large volume of spatiotemporal data has increased

[1] In the context of surface water, a watershed is a geographic area bounded peripherally by a water parting and draining to a common outlet: a point on a larger stream or river, a lake, etc.

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 in the extraction and interpretation process [8]. In this context, authors in [10] 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 because these steps have an important impact on mining results. In [1], 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 commonly used data mining algorithm can be used to extract spatial patterns. Moreover, classical data mining algorithms take a simple table as input and does not consider spatial information directly. For example, if the objective is to study changes in data generated by monitoring stations, one way of extracting such spatial patterns is to aggregate informations for each station in a single row of the input table. In [12, 18], authors use this approach, and map their spatial data to sets (or sequences) of values. Several pre-processing techniques in spatiotemporal datasets have been discussed in the literature [6, 13, 15]. Each reference has its own focus such as spatial classification, spatial association rules or knowledge discovery respectively.

Referring to **data mining**, several solutions are proposed in the literature to extract knowledge in a spatiotemporal database. Early works addressed the spatial and temporal dimensions separately. In [14], authors have studied temporal sequences which only take into account the temporal dimension. Later, in [18], authors have extended these works to represent sets of environmental features evolving over time. They extract sequences of characteristics that appear frequently in areas, but without taking into account the spatial environment. Other authors such as [11] or [16] looked for spatial patterns or co-locations, i.e., subsets of features (object-types) with instances often identified close in space. As well, in [19], authors focus on the extraction of sequences representing the propagation of spatiotemporal events in predefined time windows w.r.t. a reference location. They introduce two concepts: *Flow patterns* and *Generalized Spatiotemporal Patterns* in order to precisely extract the sequence of events that occur frequently in some locations. On the other hand, in [9], authors found that all the patterns discovered with other approaches are not all the time relevant because they may not be statistically significant and in particular not *dense* in space and time. Nevertheless, they study events one after another. Later, in [5], authors proposed the concept of *Mixed-Drove Spatiotemporal Co-occurrence Patterns*, i.e., subsets of two or more different event-types whose instances are often located in spatial and temporal proximity. But, they do not extract the frequent evolutions of even-types over time (events of each instance occur necessarily in the same time slot).

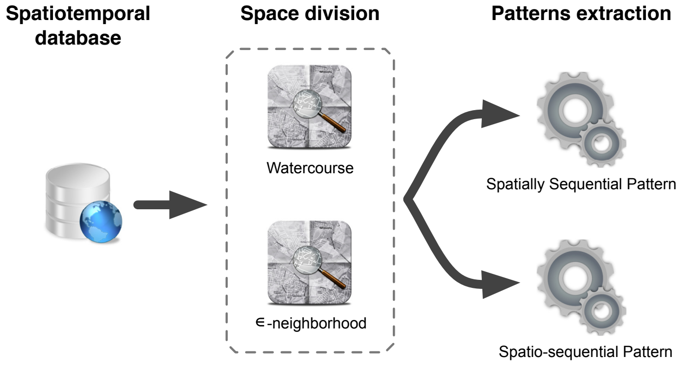

Fig. 1. Process of knowledge discovery on spatial databases

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. This first pre-processing step was combined, in the one hand, with a classical algorithm of sequential pattern mining and, in the other hand, with a new method for extracting spatiotemporal patterns (i.e., sequences of spatial sets of events). This second data mining approach allow us to deal with the developments and interactions between the study area and its immediate environment.

Several phenomena can be studied using our KDD approach, e.g., the soil erosion, the epidemic surveillance, the river pollution, and many others. In this paper, we have applied our method on data of river pollution, but our approach has been tested also on dengue epidemiological monitoring data collected in New Caledonia.

3 General process for mining spatial databases

Our approach is divided into three steps: (1) spatial decomposition and aggregation, and; (2) spatiotemporal patterns mining using two approaches. This general process is illustrated in Figure 1.

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). Thanks to this transformation, the spatial features of data are integrated into the KDD process.

The resulting spatial sequences are used as an input of the data mining step, which is composed of two methods. The first one extracts spatially frequent sequences using a classical sequential patterns mining algorithm (see, [14]), therefore, extracted patterns represent spatially frequent temporal evolutions of zones. The second one is a new approach that extracts patterns called *spatio-sequential patterns*. This last approach enables to analyze changes in zones over time taking into account their neighboring environment. Notice that *we obtain two semantically different kinds of patterns.*

3.1 Spatial pre-processing

Data at our disposal is associated to biological indicators collected by monitoring stations strategically positioned along several watersheds. This heterogeneous data is also geo-referenced and temporally variable, thus making them difficult to explore globally. Moreover, several implicit spatial relationship between studied objects may be considered. For instance, a monitoring station is located in upstream (or downstream) along to a specific watercourse, and it is also located in an agricultural zone. 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 agricultural zone, etc.). In this work, we do not study the evolution of events for each station independently, this kind of approaches are widely studied, e.g., see [18]. In this article, we propose two different ways to explore spatial data.

- A *watercourse* approach: for a given watercourse, two stations X and Y located on this watercourse are considered to be neighbors. For instance, in Figure 2, stations W, X, Y and Z belong to the same watercourse, moreover, these stations are considered as a single area and their data are combined. An example of incident that can be study 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 approach: the space is divided into areas grouping stations by exploiting the Lambert coordinates. In each of these areas, stations covering an area of ϵ km^2 are grouped, even if these stations belong to different watercourses. For instance in Figure 3, stations W, X and Y are considered as a single area, even if they are not on the same watercourse. An example of phenomenon that can be study based on 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 build *spatial sequences*, i.e., sequences containing spatial characteristics of data. Nevertheless, if we cannot perceive the "spatialization" in sequences, this feature can be evinced in patterns obtained as discussed in Section 4.1.

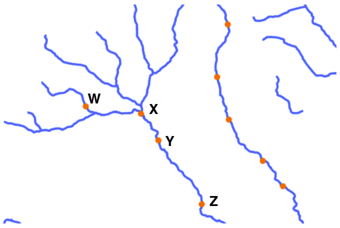

Fig. 2. *watercourse* approach

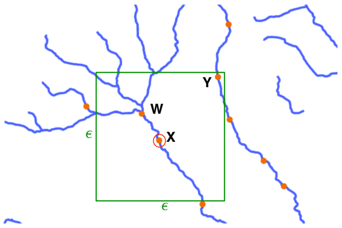

Fig. 3. *ϵ-neighborhood* approach

3.2 Data mining process

In this section, we present two data mining methods. The first one is a classical algorithm ([14]) and the second one is a new approach called **S**patio-**S**equential **P**atterns mining (or simply S2P mining).

Sequential patterns mining: Consider the spatiotemporal 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 $< IS_1, IS_2, \ldots, IS_p >$ where IS_j is an *itemset*.

A *n-sequence* is a sequence of n-itemsets. For example, consider 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 $< IS_1, IS_2, \ldots, IS_p >$ 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 $S' =< (B)(E) >$ 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.

Table 1. Example of spatiotemporal database

Client	Date	Items
Station1	2012/01/12	A, E
Station2	2012/02/28	E
Station1	2012/03/02	B, C
Station1	2012/03/12	D
Station1	2012/04/26	E

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*.

The interpretation of this first type of patterns is strongly due to how we preprocess data to be mined. Indeed, the main challenge of **spatially frequent sequences** is to capture the spatial characteristics of data grouping stations using different topologies before to data mining step. For more information, see [4].

Afterwards, we present a second type of patterns whose semantics takes into account the spatial relationships (e.g., neighborhood) between stations.

Spatio-sequential patterns mining: On the spatiotemporal database illustrated in Table 1, we define a neighborhood relationship between stations (or a set of stations), which is denoted by *Neighbor*, as:

$$\begin{cases} Neighbor(Station_i, Station_j) = true \text{ if } Station_i \text{ and } Station_j \text{ are close by} \\ Neighbor(Station_i, Station_j) = false \text{ otherwise} \end{cases}$$

We define the In relationship between *stations* and *itemsets* which describes the occurrence of itemset IS in station S at time t in the database DB: $In(IS, S, t)$ is true if *is* is present in DB for station S at time t. In our example, consider the itemset $IS = (A, E)$ then $In(IS, Station1, 2012/01/12)$ is *true* (see Table 1).

Spatial itemset. Let IS_i and IS_j be two itemsets, we say that IS_i and IS_j are spatially close if and only if $In(IS_i, Station_i, t) \wedge In(IS_j, Station_j, t) \wedge Neighbor(Station_i, Station_j)$ is *true*.

A pair of itemsets IS_i and IS_j that are spatially close, is called a **spatial itemset** and denoted by $I_{ST} = IS_i \cdot IS_j$.

To facilitate notations, we introduce a group operator for itemsets to be assigned by the operator $\cdot$ (*near*), denoted by *[]*. The θ symbol represents the *absence* of itemsets in a zone. Figure 4 shows the three types of spatial itemsets that we can build with the proposed notations. The dotted lines represent spatial neighborhood relationship.

We now define the notion of zones *evolution* according to their spatial neighborhood relationship.

Spatial Sequence. A spatial sequence or simply **S2** is an ordered list of spatial itemsets, denoted by $s = \langle I_{ST_1} I_{ST_2} \ldots I_{ST_m} \rangle$ where I_{ST_i}, $I_{ST_{i+1}}$ satisfy the constraint of temporal sequentiality . A S2 $s = \langle (AB)(\theta \cdot [B; C])(P \cdot [Q; R]) \rangle$ is illustrated in Figure 5, where the arrows represent the temporal dynamics and the dotted lines represent the environment.

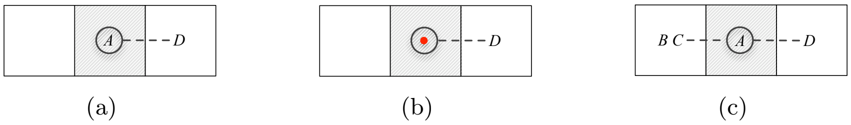

Fig. 4. Graphical representation of spatial itemsets (a) $A \cdot D$ (b) $\theta \cdot D$ (c) $A \cdot [BC; D]$

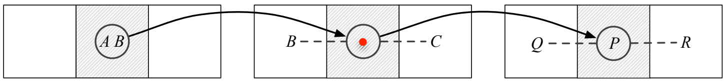

Fig. 5. Graphical representation of sequence $\langle (AB)(\theta \cdot [B; C])(P \cdot [Q; R]) \rangle$

The main challenge involved in the **spatio-sequential patterns mining problem** is to study the evolution of characteristics/events in monitoring stations taking into account immediate surrounding areas (for more information, see [2]). In providing a minimal support, the method is able to extract the frequent sequences, i.e., sequences with a support equal or greatest than a minimal support fixed by users.

More formally: Let *minsupp* be a minimum threshold set by the user, a *spatial sequence* S2 satisfying $STPi(S2) \geq minsupp$. These frequent sequences are called *spatio-sequential patterns* or simply **S2P**.

It is important to notice that both data mining methods - spatially sequential patterns and spatio-sequential patterns - can be used with any spatialization approach. For instance, monitoring station located on watercourses can be grouped by district in order to study the impact of river pollution between neighboring districts.

4 Some results

In this section, we present a qualitative evaluation by giving some examples of spatiotemporal patterns extracted by the two data mining methods. Later, these two kinds of patterns are compared semantically.

Spatially sequential patters: Table 2 shows some spatial sequential patterns extracted from *RM* water supply agency dataset using the *ϵ-neighborhood* spatialization approach. We can notice that we obtain a sequence of itemsets or set of items (events), which characterizes the evolution of a set of stations - covering 10 km^2 - over time. For instance, the spatially frequent sequence ⟨(ibd:>16.010)(ibd:(12.990;16.010] taxovar:(19.500;31.500])⟩ can be interpreted as: frequently, a high value of IBD has been register before an decrement of IDB indicator associated to a mean value of taxonomic variety.

Table 2. Some spatially sequential patterns for RM (ϵ-neighborhood) dataset for a minimal support of 0.2

Sequence	**Supp**
⟨(ibd2007:>8.121)(ibd2007:(6.810;8.121])⟩	0.33
⟨(taxovar:(19.500;31.500])(ibd:(12.990;16.010] ibd2007:(6.810;8.121])⟩	0.30
⟨(ibd:>16.010)(ibd:(12.990;16.010] taxovar:(19.500;31.500])⟩	0.24
⟨(taxovar:(19.500;31.500])(ibd:(12.990;16.010] ibd2007:(6.810;8.121] taxovar:(19.500;31.500])⟩	0.24
⟨(ibd2007:>8.121 taxovar:<=19.500 fishindex:<=8.50)⟩	0.20
...	...

Spatio-sequential patters: Table 3 shows some spatio-sequential patterns (S2P) extracted from RMC dataset using the watercourse spatialization approach. We may confirm that we obtain a sequence of itemsets (i.e., a set of items or events), which characterizes the evolution of a zone and its near surrounding over time. We should remember that a zone group a set of stations placed in a specific watercourse and stations located in close watercourses compose its near surrounding.

For instance, in Table 3, the second S2P ⟨(θ·[ibd:<=13.325; taxovar:<=15.500; ibgn:<=4.500])⟩ means that: often, a low values of IBD, taxonomic variety and IBGN indicators appear together, subsequently, we can assume that the water quality is seriously affected in some watercourses belonging the RM water supply agency. Moreover, the third S2P ⟨(ibd:>21.216)(θ·taxovar:(17.500;29.500]) (ibd:(13.985;21.216])⟩ can be interpreted by: In 30% of areas, a high value of IBD appear before the occurrence of a means value of *taxovar* in a neighbor watercourse followed by a decrement of IBD indicator.

4.1 Discusion: Spatially sequential patterns vs Spatio-sequential patterns

In this data mining process, we have focused on the extraction of spatiotemporal patterns. In this context, we have proposed two methods allowing us include spatial characteristics into the obtained patterns. These two techniques differ substantially in the process and the results are semantically different. The first one uses a widely used sequential pattern mining algorithm whereas the other uses a new method called spatio-sequential pattern mining. These two techniques have been performed on a real database that have been pre processed in order to divide the space into homogeneous zones following two pollution hypotheses (see Section 3.1).

The **first kind of patterns** represents the evolution of a set of characteristics - biological indicators - belonging to a set of monitoring stations grouped using different spatial proximities. It is important to notice that, by applying the same algorithm on the same database for the same minimal support but with the two spatial division methods, we obtain two different sets of spatially sequential patterns. This difference is reflected not only in the number of extracted patterns

Table 3. Some spatio-sequential patterns for RMC (watercourse neighborhood) dataset for a minimal support of 0.2

Sequence	**Supp**
⟨(taxovar:<=15.500)(θ· ibd:<=13.325)(θ· ibd:(13.325;20.035])⟩	0.36
⟨(θ· [ibd:<=13.325; taxovar:<=15.500; ibgn:<=4.500])⟩	0.32
⟨(ibd:>21.216)(θ· taxovar:(17.500;29.500])(ibd:(13.985;21.216])⟩	0.30
⟨(θ· ibd:(13.985;21.216])(ibd:<=13.985)(θ· ibd:<=13.985)⟩	0.27
⟨(θ· [ibd2007:<=7.510; taxovar:(15.500;29.500]])(ibd2007:(7.510;18.069])⟩	0.25
...	...

but also in their constitution themselves. To know which spatialization approach is more interesting for experts, we can apply a post processing techniques like a clustering (e.g., k-means) combined to statistical measure (e.g., sum of square for errors).

The **second kind of patterns** also represents changes in zones over time, nevertheless, it includes additional information, i.e., events appeared in neighboring areas. In this kind of extracted patterns, we can directly perceive the spatial relationships between neighboring areas thanks to spatial operator • (close to). The extraction of this additional information impacts directly in the performance of our algorithm since the search space increases with the number of neighbors to be evaluated. In contrast, this additional information can be crucial in decisions concerning the preservation and restoration of rivers and their surrounding environments.

It is important to notice that, the first kind of patterns are included in the second one. Indeed, the spatio-sequential pattern mining is an extension of the spatially sequential patterns mining taking into account the neighboring areas.

These two proposed approaches are generic. Indeed, we have also applied to our approaches to other real datasets, e.g., some results for epidemic monitoring of dengue fever and a visualization prototype are available on `http://datamining.univ-nc.nc/`.

5 Conclusion and perspectives

In this paper we have presented the first steps of a data mining project on hydrological data. In particular, we applied two algorithms for spatiotemporal pattern extraction according to two spatialization approaches. Moreover, a detailed comparison between these two data mining techniques has been included in this work. We highlighted the problems that are posed depending on choices made in terms of spatialization and their influence on the number of extracted 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. Then, for the extraction phase, we would like to compare different data mining techniques in terms of obtained patterns. In addition, a huge number of patterns have been extracted. Currently, we have proposed a new quality measure called *the least temporal contradiction* to filter the most relevant patterns. This measure allow us to estimate how many times a rule is verified *vs* how many times it is disabled. A pattern that is most frequently tested as disabled is a priori irrelevant. This measure is being adapted to spatio-sequential patterns.

We also have proposed a visualization prototype, which is available on `http://datamining.univ-nc.nc/` and allow us to visualize a spatial dynamic of spatio-sequential patterns extracted on dengue fever dataset.

References

1. Bogorny, V., Engel, P., Alvares, L. O., *Spatial data preparation for knowledge discovery.* IEEE Computer Graphics pp. 24 (5), 8 (2005).
2. Alatrista-Salas, H., Bringay, S., Flouvat, F., Selmaoui-Folcher, N. and Teisseire, M. *The Pattern Next Door: Towards Spatio-sequential Pattern Discovery.* Advances in Knowledge Discovery and Data Mining - 16th Pacific-Asia Conference, PAKDD, pp. 157-168 (2012)
3. Cao, L., Zhang, H., Zhao, Y., Luo, D., Zhang, C., *Combined mining: Discovering informative knowledge in complex data.* Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 41 (3), pp. 699–712 (2011).
4. Alatrista-Salas, H., Cernesson, F., Bringay, S., Azé, J., S., Flouvat, F., Selmaoui-Folcher, N. and Teisseire, M. *Recherche de séquences spatio-temporelles peu contredites dans des données hydrologiques.* Revue des Nouvelles Technologies de l'Information (RNTI), RNTI-E-22, pp. 165–188 (2011)
5. Celik, M., Shekhar, S., Rogers, J., and Shine, J., *Mixed-drove spatiotemporal co-occurrence pattern mining.* Proc. of IEEE TKDE, 20(10), pp. 1322–1335 (2008).
6. Elias, B., *Extracting landmarks with data mining methods.* In: Spatial Information Theory. Foundations of Geographic Information Science. Vol. 2825 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 375–389 (2003).
7. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., *Advances in knowledge discovery and data mining.* American Association for Artificial Intelligence, Menlo Park, CA, USA, Ch. From data mining to knowledge discovery: an overview, pp. 1–34 (1996).
8. Hsu, W., Lee, M., Wang, J., *Temporal and Spatio-Temporal Data Mining.* Gale Virtual Reference Library. IGI Pub (2008).
9. Huang, Y., Zhang, L., and Zhang, P., *A framework for mining sequential patterns from spatio-temporal event data sets.* Proc. of IEEE TKDE, 20(4) pp. 433–448 (2008).
10. Gibert, K., Izquierdo, J., Holmes, G., Athanasiadis, I., Comas, J., Sanchez-Marre., *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, pp. 1937–1958 (2008).
11. Han, J., Koperski, K., and Stefanovic, N., *Geominer: a system prototype for spatial data mining.* In Proc. of ACM SIGMOD, SIGMOD '97, pp. 553–556 (1997).

12. Koperski, K., Han, J., *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, pp. 47–66 (1995).
13. Mennis, J., Liu, J. W., *Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change.* Transactions in GIS 9 (1), pp. 5–17 (2005).
14. Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., Hsu, M.-C., *Mining sequential patterns by pattern-growth: The prefixspan approach.* IEEE Transactions on Knowledge and Data Engineering 16 (11), pp. 1424–1440 (2004).
15. Qi, F., Zhu, A., *Knowledge discovery from soil maps using inductive learning.* International Journal of Geographical Information Science 17 (8), pp. 771–795 (2003).
16. Shekhar, S. and Huang, Y., *Discovering Spatial Co-Location Patterns A Summary Of Results.* Advances in Spatial and Temporal Databases, pages pp. 236–256 (2001).
17. Triki, D., Frihida, A., Ben Ghezala, H., Claramunt, C., *Modèle et langage pour la manipulation de trajectoires spatio-temporelles.* In: International Journal of Geomatics and Spatial Analysis IJGSA, vol. 20/1, pp. 37–64 (2010).
18. Tsoukatos, I., Gunopulos, D., *Efficient mining of spatiotemporal patterns. In: Advances in Spatial and Temporal Databases.* In Lecture Notes in Computer Science. Springer Berlin / Heidelberg, Vol. 2121, pp. 425–442 (2001).
19. Wang, J., Hsu, W., and Lee, M., *Mining generalized spatio-temporal patterns.* In Database Systems for Advanced Applications, Springer, pp. 649–661 (2005).