The Pattern Next Door: Towards Spatio-sequential Pattern Discovery

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

The Pattern Next Door: Towards Spatio-Sequential Pattern Discovery

Hugo Alatrista Salas[1,3], Sandra Bringay[2], Frédéric Flouvat[3], Nazha Selmaoui-Folcher[3], and Maguelonne Teisseire[1,2]

[1] IRSTEA, UMR TETIS, 500 rue Jean-François Breton, 34093 Montpellier - France
firstname.lastname@teledetection.fr
[2] LIRMM, UMR 5506, 161 rue Ada, 34392 Montpellier - France
firstname.lastname@lirmm.fr
[3] PPME, Université de la Nouvelle-Calédonie, BP R4, Nouméa - New Caledonia
firstname.lastname@univ-nc.nc

Abstract. Health risks management such as epidemics study produces large quantity of spatio-temporal data. The development of new methods able to manage such specific characteristics becomes crucial. To tackle this problem, we define a theoretical framework for extracting spatio-temporal patterns (sequences representing evolution of locations and their neighborhoods over time). Classical frequency support doesn't consider the pattern neighbor neither its evolution over time. We thus propose a new interestingness measure taking into account both spatial and temporal aspects. An algorithm based on pattern-growth approach with efficient successive projections over the database is proposed. Experiments conducted on real datasets highlight the relevance of our method.

1 Introduction

In everyday life, we can observe many phenomena occurring in space and time simultaneously. For example, the movements of a person associate spatial information (e.g. the departure and arrival coordinates) and temporal information (e.g. the departure and arrival dates). Other applications, with more complex dynamics, are much more difficult to analyze. It is the case of spread of infectious disease, which associates spatial and temporal information such as the number of patients, environmental or entomological data. Yuang in [13] describes this concept of dynamics as a *set of dynamic forces impacting the behavior of a system and components, individually and collectively.*

In this paper, we focus on spatio-temporal data mining methods to better understand the dynamics of complex systems for epidemiological surveillance. In the case of dengue epidemics, public health experts know that the evolution of the disease depends on environmental factors (e.g. climate, areas with water points, mangroves...) and interactions between human and vector transmission (e.g. the mosquito that carries the disease). However, the impact of environmental factors and their interactions remain unclear.

To address these issues, spatio-temporal data mining provides highly relevant solutions through the identification of relationships among variables and events, characterized in space and time without *a priori hypothesis*. For example, in our context, we will discover combinations of changes in environmental factors that lead epidemic peaks in specific spatial configurations. We will show in the related works section that existing methods are not completely adapted to our problem. For this reason, we have define new spatio-sequential patterns, based on an extension of sequential patterns, to link the spatial and temporal dimension. An example of pattern in the dengue context is: *frequently over the past 10 years, if it rains in an area and if there is standing water and high temperatures in the neighborhood, then there is an increase number of mosquitoes in adjacent areas, followed by an increase of dengue cases*. It can be used for analysis by health care professionals, to better understand how environmental factors influence the development of epidemics. Such patterns are very interesting because they enable to capture evolution of areas considering their events and events in adjacent zones. However, they are very difficult to mine because the search space is very large. Proposing scalable methods to find these patterns are consequently very challenging. We have defined an interestingness measure to overcome this problem of scalability and an efficient algorithm based on pattern-growth approach.

In section 2, we review existing spatio-temporal data mining methods and we show that these methods are not suitable for our problem. In section 3, we detail our theoretical framework. In section 4, we present our algorithm called *DFS-S2PMiner*. In section 5, we present experiments on real datasets. The paper ends with our conclusions and future perspectives.

2 Related work

In this related work section, we are not concerned by the trajectories problematic addressed in [1, 3]. We only focus on methods analyzing the evolution and the interaction of objects or events characteristics through space and time. Early work addressed the spatial and temporal dimensions separately. For example, Han et al. in [4] or Shekhar et al. in [10] looked for spatial patterns or co-location, i.e. subsets of features (object-types) with instances often identified as close in space. In our context, an example of co-location is *within a radius of 200 m, mosquitoes nests are frequently found near ponds.* On the contrary, other authors as Pei et al. in [9] have studied temporal sequences which only take into account the temporal dimension. Tsoukatos et al. in [11] have extended these works to represent sets of environmental features evolving in time. They extract sequences of characteristics that appear frequently in areas, but without taking into account the spatial neighborhood. An example of pattern obtained is: *in many areas, heavy rain occurs before the formation of a pond, followed by the development of mosquito nest*. If these two types of methods, only spatial or temporal, can be very relevant for epidemiological surveillance, they do not capture relations such as: *often, a heavy rain occurs before the formation of a pond followed in a close area by the development of mosquito nests*. In [12], Wang et al. focus on the

extraction of sequences representing the propagation of spatiotemporal events in predefined time windows. They introduce two concepts: *Flow patterns* and *Generalized Spatiotemporal Patterns* in order to extract precisely the sequence of events that occur frequently in some locations. Thus, the authors will be able to identify patterns of the form: *dengue cases appear frequently in area Z1 after the occurrence of high temperatures and the presence of ponds in area Z2.*

However, Huang et al. in [7] found that all the patterns discovered with others approaches are not all the time relevant because they may not be statistically significant and in particular not *"dense"* in space and time. They therefore proposed an interestingness measure taking into account the spatial and temporal aspects to extract global sequence of features. However, they study the events one after another. They don't take into account the interactions such as *often heavy rain and the occurrence of ponds are presented before the development of mosquito nests.* Celik et al. in [2], 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 (*e.g.* an event-type is *heavy rain* and an instance is *heavy rain in zone Z1 the 10/17/2011*). For similar reasons than Huang, they have proposed a specific monotonic composite interest measure based on spatial and temporal prevalence measures. However, they do not extract the frequent evolutions of event-types over time (events of each instance occur necessarily in the same time slot). For example, we can only extract patterns such as: *heavy rain, ponds and development of mosquito nests are frequently found together in lots of time slots.* Finally, approaches proposed by Wang, Huang and Celik cannot capture the evolution of areas with regard to their set of event-types and the sets of event-types of their neighbors.

In this paper, we describe a method for extracting spatio-temporal sequences of patterns (i.e. sequences of spatial sets of events) called *S2P* (Spatio-Sequential Patterns). We aim at identifying relationships such as: *the presence of dengue cases in an area is often preceded of high temperatures and the presence of water tanks in a neighboring area.* Thus, we will deal with the developments and interactions between the study area and its immediate environment. Moreover, as this type of patterns are very difficult to mine, because of the huge generated search space, we will introduce an interestingness measure to make our approach scalable.

3 Spatio-Sequential Patterns: Concepts and Definitions

3.1 Preliminaries

A spatio-temporal database is a structured set of information including geographic components (e.g. neighborhoods, rivers, etc.) and temporal components (e.g. rain, wind). Such a database is defined as a triplet $DB = (D_{\mathrm{T}}, D_{\mathrm{S}}, D_{\mathrm{A}})$ where D_{T} is the temporal dimension, D_{S} the spatial dimension and $D_{\mathrm{A}} = \{D_{\mathrm{A}_1}, D_{\mathrm{A}_2}, \ldots, D_{\mathrm{A}_p}\}$ a set of analysis dimensions associated with attributes. The *temporal dimension* is associated with a domain of values denoted $dom(D_{\mathrm{T}}) = \{T_1, T_2, \ldots, T_t\}$ where $\forall i \in [1..t]$, T_i is a *timestamp* and $T_1 < T_2 < \ldots < T_t$. The

spatial dimension is associated with a domain of values denoted $dom(D_{\mathrm{S}}) = \{Z_1, Z_2, \ldots Z_l\}$ where $\forall i \in [1..l]$, Z_i is a *zone*. We define on $dom(D_{\mathrm{S}})$ a neighborhood relationship, denoted *Neighbor* by:

$$Neighbor(Z_i, Z_j) = true \text{ if } Z_i \text{ and } Z_j \text{ are neighbors, } false \text{ otherwise} \quad (1)$$

Each dimension D_{A_i} ($\forall i \in [1..p]$) in the set of *analysis dimensions* D_{A}, is associated with a domain of values denoted $dom(A_i)$. In these domains, the values can be ordered or not.

To illustrate the definitions, we use a sample of weather database, Table 1, which represents weather in three cities on three consecutive days. The table lists temperature (Temp), precipitation (Prec), wind speed (Wind) and gusts in Km/h. The three cities are associated by a neighborhood relationship described in Figure 1.

Table 1: Weather changes in three cities : Z_1, Z_2 et Z_3 on December 22, 23, 24, 2010.

City	Date	Temp	Prec	Wind	Gusts
Z_1	12/22/10	T_m	P_m	V_m	-
Z_1	12/23/10	T_m	P_m	V_l	-
Z_1	12/24/10	T_l	P_m	V_m	55
Z_2	12/22/10	T_m	P_m	V_m	-
Z_2	12/23/10	T_l	P_m	V_l	-
Z_2	12/24/10	T_l	P_l	V_m	-
Z_3	12/22/10	T_l	P_m	V_s	75
Z_3	12/23/10	T_m	P_s	V_l	-
Z_3	12/24/10	T_l	P_s	V_s	55

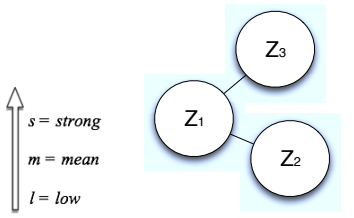


Fig. 1: Neighboring cities.

In Table 1, $D_{\mathrm{T}} = \{Date\}$, $D_{\mathrm{S}} = \{City\}$ and $D_{\mathrm{A}} = \{Temp, Prec, Wind, Gusts\}$. The domain of the temporal dimension is $dom(D_{\mathrm{T}}) = \{12/22/10, 12/23/10, 12/24/10\}$ with $12/22/10 < 12/23/10 < 12/24/10$. The domain of spatial dimension is $dom(D_{\mathrm{S}}) = \{Z_1, Z_2, Z_3\}$ with $Neighbor(Z_1, Z_2) = true$, $Neighbor(Z_1, Z_3) = true$ and $Neighbor(Z_2, Z_3) = false$. Finally, for the analysis dimensions *Temp* and *Gusts*, the domains are respectively $dom(Temp) = \{T_m, T_l, T_s\}$ and $dom(Gusts) = \{55, 75\}$.

3.2 Spatio-sequential patterns

Definition 1. ***Item and Itemset.*** *Let I be an* item*, a literal value for the dimension D_{A_i}, $I \in dom(D_{A_i})$. An* itemset*, $IS = (I_1 I_2 \ldots I_n)$ with $n \leq p$, is a non empty set of* items *such that $\forall i, j \in [1..n], \exists k, k' \in [1..p]$, $I_i \in dom(D_{A_k})$, $I_j \in dom(A_{D_{A'_k}})$ and $k \neq k'$.*

All items in an itemset are associated with different dimensions. An itemset with k items is called k-itemset.

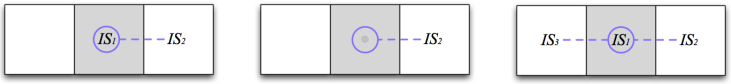
We define the In relationship between *zones* and *itemsets* which describes the occurrence of itemset IS in zone Z at time t in the database DB:

$In(IS, Z, t)$ is true if IS is present in DB for zone Z at time t. In our example, consider the itemset $IS = (T_m P_m V_l)$ then $In(IS, Z_1, 12/23/10)$ is *true* as the itemset $(T_m P_m V_l)$ occurs for zone Z_1 on 12/23/10 (see Table 1).

We now define the notion of *interaction* with neighbor zones.

Definition 2. ***Spatial itemset.*** *Let IS_i and IS_j be two itemsets, we say that IS_i and IS_j are spatially close iff $\exists Z_i, Z_j \in dom(D_S)$, $\exists t \in dom(D_T)$ such that $In(IS_i, Z_i, t) \wedge In(IS_j, Z_j, t) \wedge Neighbor(Z_i, Z_j)$ 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 *n-ary* group operator for itemsets to be assigned by the operator $\cdot$ (*near*), denoted []. The θ symbol represents the *absence* of itemsets in a zone. Figure 2 shows the three types of spatial itemsets that we can build with the proposed notations. The dotted lines represent the spatial dynamics.



(a) (b) (c)

Fig. 2: Graphical representation of spatial itemsets (a) $IS_1 \cdot IS_2$ (b) $\theta \cdot IS_2$ (c) $IS_1 \cdot [IS_2; IS_3]$.

The spatial itemset $I_{ST} = (T_m \cdot (V_l P_m))$ describes that events T_m and $V_l P_m$ occur in neighboring zones at the same time. The spatial itemset $I_{ST} = (\theta \cdot [T_m; P_l])$ indicates that T_m and P_l occur in two different zones neighbor to a zone where no event appears.

Definition 3. ***Inclusion of spatial itemset.*** *A spatial itemset $I_{ST} = IS_i \cdot IS_j$ is* included*, denoted $\subseteq$, in another spatial itemset $I'_{ST} = IS'_k \cdot IS'_l$, iff $IS_i \subseteq IS'_k$ and $IS_j \subseteq IS'_l$.*

The spatial itemset $I_{ST} = (T_m P_m \cdot V_l)$ is *included* in the spatial itemset $I'_{ST} = (T_m P_m \cdot V_l 55)$ because $(T_m P_m) \subseteq (T_m P_m)$ and $(V_l) \subseteq (V_l, 55)$.

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

Definition 4. ***Spatial Sequence.*** *A spatial sequence or simply* ***S2*** *is an ordered list of spatial itemsets, denoted $s = \langle I_{ST_1} I_{ST_2} \ldots I_{ST_m} \rangle$ where I_{ST_i}, $I_{ST_{i+1}}$ satisfy the constraint of temporal sequentiality for all $i \in [1..m-1]$.*

A S2 $s = \langle (T_m)(\theta \cdot [P_l; V_s])(V_l \cdot [P_l; T_l]) \rangle$ is illustrated in figure 3 for the zone $Z1$, where the arrows represent the temporal dynamics and the dotted lines represent the environment.

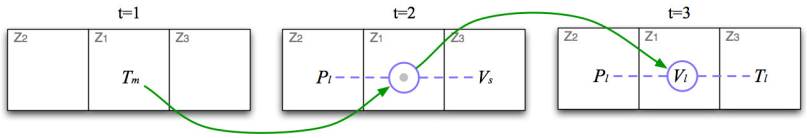


Fig. 3: Example of the spatio-temporal dynamic.

A relationship generalization (or specialization) between S2's is defined as follows:

Definition 5. ***Inclusion of S2.*** *A S2* $s = \langle I_{ST_1} I_{ST_2} \dots I_{ST_m} \rangle$ *is more specific than a S2* $s' = \langle I'_{ST_1} I'_{ST_2} \dots I'_{ST_n} \rangle$*, denoted* $s \preceq s'$*, if there exists* $j_1 \leq \dots \leq j_m$ *such that* $I_{ST_1} \subseteq I'_{ST_{j_1}}, I_{ST_2} \subseteq I'_{ST_{j_2}}, \dots, I_{ST_m} \subseteq I'_{ST_{j_m}}$.

A S2 $s = \langle (T_l P_m \cdot P_l V_s)(55) \rangle$ is included in the S2 $s' = \langle (T_l P_m \cdot P_l V_s)(55 \cdot V_s) \rangle$ because $(T_l P_m \cdot P_l V_s) \subseteq (T_l P_m \cdot P_l V_s)$ and $(55) \subseteq (55 \cdot V_s)$.

For a specific zone, we note s_Z the associated spatial data sequence in the database DB. s_Z contains or supports a spatial sequence s if s is a subsequence of s_Z. The support of a spatial sequence s is thus defined as the number of zone supporting s. If the support of the spatial sequence is greater than a user-defined threshold, the sequence is frequent and corresponds to a **spatio-sequential pattern** ($S2P$). Nevertheless, in a spatio-temporal context, we need to define a more precise and suitable prevalence measure, as explained in the next section.

3.3 Spatio-temporal participation

The proposed spatio-sequential pattern allow to tackle both spatial and temporal issues. In order to manage in an efficient way the mining of such patterns, a new filtering measure has to be defined. To highlight the participation of an item in a spatial sequence, we propose an adaptation of the participation index [6] which is a combination of two measures: **spatial participation index** and **temporal participation index** taking into account respectively the spatial dimension and the number of occurrences in time.

Definition 6. ***Spatial participation ratio*** *Let* s *be a spatial sequence and* I *be an item of* s*, the* spatial participation ratio *for* I *in* s*, denoted by* $SPr(s, I)$ *is the number of zones which contain* s *divided by the number of zones where the item* I *appears in the whole database:*
$SPr(s, I) = \frac{Supp(s)}{Supp(I)}$

Definition 7. ***Spatial participation index*** *Let* $s = \langle I_{ST_1}, I_{ST_2}, \dots I_{ST_n} \rangle$ *be a spatial sequence, the* spatial participation index *of* s *denoted* $SPi(s)$ *is the minimum of* spatial participation ratio*:*
$SPi(s) = MIN_{\forall I \in dom(A), I \in s} \{SPr(s, I)\}$

Definition 8. ***Temporal participation ratio*** *Let s be a spatial sequence and I be an item of s, the* temporal participation ratio *for I in s denoted $TPr(s, I)$ is the number of occurrences of s (i.e. the number of instances over time) divided by the total number of occurrences of I:*
$TPr(s, I) = \frac{NbOccurrences(s)}{NbOccurrences(I)}$

Definition 9. ***Temporal participation index*** *Let $s = \langle I_{ST_1}, I_{ST_2}, \dots I_{ST_n} \rangle$ be a spatial sequence, the* temporal participation index *of s denoted $TPi(s)$ is the minimum of* temporal participation ratio*:*
$TPi(s) = MIN_{\forall I \in dom(A_i), I \in s} \{TPr(I, s)\}$

We define the **spatio-temporal participation index** of a spatial sequence s, $STPi(s)$, as:

$$STPi(s) = 2 * \frac{SPi(s) * TPi(s)}{SPi(s) + TPi(s)} \tag{2}$$

Given a spatio-temporal database DB, the problem of spatio-sequential pattern mining is to find all spatial sequences whose spatio-temporal participation index is greater than a user-specified threshold *min_stpi*.

Note that the predicate "STPi is greater than a user-threshold" is antimonotonic. If a spatio-sequential pattern s is not frequent, all patterns s' such as s is included in s' ($s \preceq s'$), are also not frequent. This property is used in our pattern mining algorithm to prune the search space and quickly find frequent spatio-sequential patterns.

4 Extraction of spatio-sequential patterns

In this section, we propose an algorithm called *DFS-S2PMiner* to extract spatio-sequential patterns considering both spatial and temporal aspects. DFS-S2PMiner adopts a depth-first-search strategy based on successive projections of the database such as FP-Growth [5] and Prefixspan [8] for scalability purpose. Specifically, this algorithm is based on the *pattern-growth* strategy used in [5]. The principle of this approach is to extract frequent patterns without a candidate generation step. This approach recursively creates a projected database, associates it with a fragment of frequent pattern, and "mines" each projected database separately. The frequent patterns are extended progressively along a depth-first exploration of the search space.

First, we introduce the definition of the projection of a spatio-temporal database used in the algorithm. Let s be a spatio-sequence of the database DB. The projection of database DB w.r.t. s, denoted $DB|_s$, is the set of suffixes of s in DB.

The algorithm 1 describes our recursive algorithm DFS-S2PMiner. First, the set of frequent items I and $\theta \cdot I$, denoted F_1, is extracted from the projected database $DB|_\alpha$ (line 1 of Algorithm 1). These items constitute extensions of sequence α. Note that in the first recursive call, $DB|_\alpha$ corresponds to the initial database DB (since $\alpha = \{\}$). Then, for each of these items $X \in F_1$, we extend

the spatio-sequential pattern α with X (lines 3 and 4). Two types of extension are possible : 1) adding X to the last spatial itemset of the sequence α (line 3) or 2) inserting X after (i.e. the next time) the last spatial itemset of α (line 4). We check the measure of interest for these two spatio-sequential patterns (lines 5 and 9) and record frequent ones in the set of solutions F (lines 6 and 10). For each frequent pattern, the algorithm then performs another projection of the database using $DB|_\alpha$ and recursively extends the pattern by invoking again the algorithm (lines 7 and 11). The algorithm stops when no more projections can be generated.

Algorithm 1 DFS-S2PMiner

– **Main routine**

Require: A spatio-temporal database DB and a user-defined threshold min_stpi
Ensure: A set of frequent spatio-sequential patterns F
$\alpha \leftarrow \{\}$
Call $Prefix\text{-}growthST(\alpha, min\_stpi, DB|_\alpha, F)$

– **Prefix-growthST** $(\alpha, min\_stpi, DB|_\alpha, F)$

Require: a spatio-sequential pattern α , the user-defined threshold min_stpi, the projection $DB|_\alpha$ of the spatio-temporal database on α, and F a set of frequent spatio-temporal patterns;
1. $F_1 \leftarrow$ {a set of frequents items I and $\theta \cdot I$ on $DB|_\alpha$, with $I \in \bigcup_{i \in [1..p]} dom(D_{A_i})$ }
2. **for all** $X \in F_1$ **do**
3. $\beta \leftarrow \alpha X$
4. $\delta \leftarrow \alpha(X)$
5. **if** $STPi(\beta) \geq min\_stpi$ **then**
6. $F \leftarrow F \cup \beta$;
7. Prefix-growthST$(\beta, min_stpi, DB|_\beta, F)$
8. **end if**
9. **if** $STPi(\delta) \geq min\_stpi$ **then**
10. $F \leftarrow F \cup \delta$;
11. Prefix-growthST$(\delta, min_stpi, DB|_\delta, F)$
12. **end if**
13. **end for**

We use our running example (Table 1 and Figure 1) with $min\_stpi = 2/3$ to illustrate this algorithm.

Iteration 1 ($\alpha = \{\}$)

– ***Extraction on frequent items and spatial items (line 1).*** The first step is to extract frequent items and frequent spatial items from DB, let:

$$F_1 = \{P_m : 3, T_m : 3, V_m : 2, V_l : 3, T_l : 3, 55 : 2, \theta \cdot T_m : 3, \\ \theta \cdot P_m : 3, \theta \cdot V_m : 3, \theta \cdot V_l : 3, \theta \cdot T_l : 3, \theta \cdot 55 : 3\}$$

– ***Extension of current sequence α (lines 3-4).***
– ***STPi processing and Recording solutions (lines 5-6 and 9-10).***
– ***Projection and Recursive call (lines 7 and 11).*** For each frequent item I and $\theta \cdot I$, the algorithm calculates the corresponding projection of the database. For example, for the frequent item P_m, we obtain the following projection (see Table 2). Each of these projected database is used in a recursive call to find its frequent super-sequences.

Table 2: Projected database of $\langle(P_m)\rangle$

Zones	Sequences	Neighbors	Neighbor sequences
Z_1	$S_1 = \langle(\_\boldsymbol{V_m})(T_m\boldsymbol{P_m}V_l)(T_l\boldsymbol{P_m}\boldsymbol{V_m}55)\rangle$	Z_2	$S_2 = \langle(\_V_m)(T_lP_mV_l)(T_lP_lV_m)\rangle$
		Z_3	$S_3 = \langle(\_V_s75)(T_mP_sV_l)(T_lP_sV_s55)\rangle$
Z_2	$S_2 = \langle(\_\boldsymbol{V_m})(T_l\boldsymbol{P_m}V_l)(T_lP_l\boldsymbol{V_m})\rangle$	Z_1	$S_1 = \langle(\_V_m)(T_mP_mV_l)(T_lP_mV_m55)\rangle$
Z_3	$S_3 = \langle(\_V_s75)(T_mP_sV_l)(T_lP_sV_s55)\rangle$	Z_1	$S_1 = \langle(\_V_m)(T_mP_mV_l)(T_lP_mV_m55)\rangle$

Iteration 2 ($\alpha = \langle(P_m)\rangle$)

- ***Extraction on frequent items and spatial items (line 1).*** The first recursive call will build the super-sequences with the prefix $\langle(P_m)\rangle$ from the projected database of Table 2. Specifically, the algorithm will find frequent items in the projected database (line 1) and extend $\langle(P_m)\rangle$ (line 2 - 4). The frequent items obtained from $DB|_{\langle(P_m)\rangle}$ are: $\{V_m : 2, T_m : 2, P_m : 2, V_l : 3, T_l : 3, 55 : 2, \theta \cdot V_m : 3, \theta \cdot T_l : 3, \theta \cdot P_m : 3, \theta \cdot V_l : 3, \theta \cdot T_m : 3, \theta \cdot 55 : 3\}$
- ***Extension of current sequence α (lines 3-4).*** The first frequent item found is $\langle V_m\rangle : 2$. Therefore, we can build two spatial sequences: $\langle(P_mV_m)\rangle$ (line 3) and $\langle(P_m)(V_m)\rangle$ (line 4).
- ***STPi processing and Recording solutions (lines 5-6 and 9-10).*** The spatio-sequential pattern $\langle(P_m)(V_m)\rangle$ with $STPi = 2/3$ is frequent (line 9).
- ***Projection and Recursive call (lines 7 and 11).*** Thus, the algorithm uses this pattern to make a new projection (see Table 3) and to recursively search all frequent super-sequences with the prefix $\langle(P_m)(V_m)\rangle$.

Table 3: Projected database of $\langle(P_m)(V_m)\rangle$

Zones	Sequences	Neighbors	Neighbor sequences
Z_1	$S_1 = \langle(T_mP_mV_l)(T_lP_mV_m55)\rangle$	Z_2	$S_2 = \langle(T_l\boldsymbol{P_m}V_l)(T_l\boldsymbol{P_m}V_m)\rangle$
		Z_3	$S_3 = -$
Z_2	$S_2 = \langle(T_lP_mV_l)(T_lP_lV_m)\rangle$	Z_1	$S_1 = \langle(T_m\boldsymbol{P_m}V_l)(T_l\boldsymbol{P_m}V_m55)\rangle$
Z_3	$S_3 = \emptyset$	Z_1	$S_1 = \langle(T_m\boldsymbol{P_m}V_l)(T_l\boldsymbol{P_m}V_m55)\rangle$

Iteration 3 ($\alpha = \langle(P_mV_m)\rangle$)

- ***Extraction on frequent items and spatial items (line 1).*** The frequent items obtained for $DB|_{\langle(P_mV_m)\rangle}$ are: $\{V_m : 2, P_m : 2, V_l : 2, T_l : 2, \theta \cdot V_m : 3, \theta \cdot T_l : 3, \theta \cdot P_m : 3, \theta \cdot V_l : 3, \theta \cdot 55 : 3\}$.
- ***Extension of current sequence α (lines 3-4).*** For example, the spatial item $\theta \cdot P_m : 3$ is one of the frequent items. In this case, the algorithm builds the spatio-sequential pattern $\langle(P_m)(V_m)(\theta \cdot P_m)\rangle$.
- ***STPi processing and Recording solutions (lines 5-6 and 9-10).*** This pattern is frequent with a $STPi = 1$ because $\langle\theta \cdot P_m\rangle$ appears in all times and zones (see Table 3).

– ***Projection and Recursive call (lines 7 and 11).*** When all frequent items are projected, the algorithm goes through another *branch* of the search space, i.e. patterns beginning with $\langle(T_m)\rangle$ (see set F_1)

The algorithm thus proceeds generally in the same way whether items are spatial or not. The main difference is how to compute the support. The support of a spatial item is the number of zones where the item occurs at least once in their neighborhood (so we have $\theta \cdot V_l : 3$ in Table 2). Notice that when the algorithm extends a pattern of type $\langle(I_{ST_1})(I_{ST_2})...(I_{ST_k} \cdot X)\rangle$ with a common item $\theta \cdot Y$, the operator of *n-ary* group is used to represent the sequence as $\langle(I_{ST_1})(I_{ST_2}) \ldots (I_{ST_k} \cdot [X;Y])\rangle$.

5 Experiments

The approach proposed in this paper has been integrated in a Java prototype, and it has been experimented on two real datasets. The first one represents the evolution of dengue infection in a city during an epidemic (26 dates). The city is divided in 81 districts each one characterized by 12 epidemic and environmental attributes (e.g. number of dengue cases, precipitation per day or presence of pools). The second dataset is a record of biological indicators in the Saône rivers, for example, IBGN (Standardized Global Biological Index) and IBD (Biological Diatom Index). These indicators are associated with hydrological stations along the watercourse and raised up made by some stations along the watersheds of the Saône. This dataset includes 815 samples associated to 223 stations (zones) and 10 attributes.

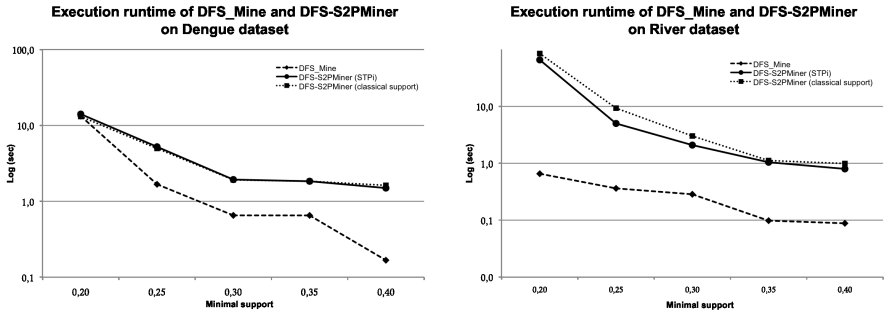
We compared our approach with the work proposed by Tsoukatos [11] since it is the closest work. Indeed, this work extracts sequences of itemsets representing the evolution of each zone individually (but without taking into account neighbors as in our approach). Experiments have been done on an Intel Core I5 processor with 4G of RAM on Linux.

First, a qualitative evaluation of the results was done. We compared the patterns obtained by our approach with the ones obtained by the DFS_Mine algorithm of Tsoukatos on the dengue dataset.

For example, both approaches could find classical sequential patterns such as *"few pools, few precipitations and few graveyard are followed by few dengue, few precipitations and wind"*. However, our approach could also find complex patterns such as *"few pools, few precipitations and few graveyard, followed by few pools and few precipitations in neighbor zones, are followed by few dengue in neighbor zones"*. This example gives an idea of the richness of our patterns by enabling to highlight the influence of neighbor areas.

When using the spatio-temporal participation index as measure of interest, we can't compare any further the extracted patterns since prevalence measures are different. While the approach of Tsoukatos keeps sequences occurring in many zones but not necessarily several times, our approach keeps sequences occurring in many zones and several times. The interest of our proposal is to consider the temporal weight of patterns.

Second, a quantitative evaluation of our approach was done. We compared the execution time of our algorithm with the DFS_Mine algorithm proposed by Tsoukatos in [11]. Figure 4 shows execution times of DFS_Mine (classical support) and DFS-S2PMiner algorithms (using classical support and spatio-temporal participation index) on the studied datasets for several thresholds. Execution times are relatively similar while our approach is doing more complex processing. Indeed, as shown by figure 5, the STPi measure allows an efficient pruning of the search space, even for the large dataset of the Saône river.



(a) (b)

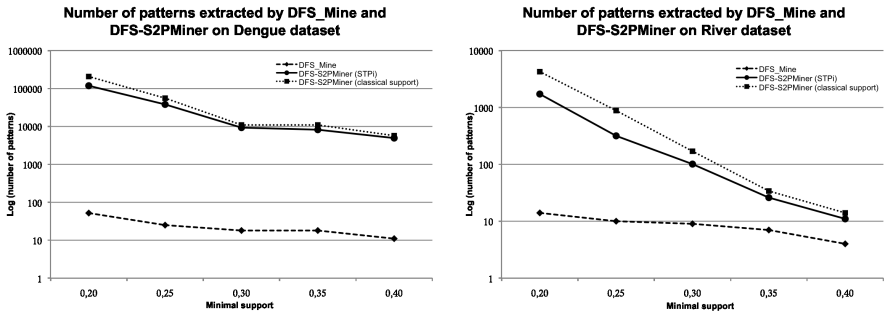
Fig. 4: Execution runtime of DFS_Mine and DFS-S2PMiner algorithms on (a) Dengue dataset (b) River dataset

6 Conclusion & Perspectives

In this paper, we propose a new concept of spatio-temporal patterns called spatio-sequential patterns *(S2P)*. This concept enables to analyze the evolution of areas considering their set of features and their neighboring environment. An example application of these patterns is the study of the spatiotemporal spread of dengue w.r.t. epidemic, district and environmental data. A formal framework is established to define *S2P* generically. To extract these patterns, we propose a generic method called *DFS-S2PMiner* based on a depth-first strategy. A new prevalence measure has been defined to cope with the limits of the classical support w.r.t. spatial and temporal aspects. Our proposal has been experimented on two real datasets. Results show the interest of the approach to extract efficiently rich spatio-temporal patterns.

Among possible future developments, we plan to extend the concept of neighborhood to n neighborhoods while allowing scalability. No new definitions are needed but an heuristic exploration of the search space may be required.

Acknowledgments. We wish to thank the Department of Health and Social Affairs of New Caledonia, The Institute Pasteur, IRD and UNC for giving us the `Dengue` data set (Convention 2010). This work was partly funded by French contract ANR-2010-COSI-012 FOSTER.



(a) (b)

Fig. 5: Number of patters extracted by DFS_Mine and DFS-S2PMiner algorithms on (a) Dengue dataset (b) River dataset

References

[1] H. Cao, N. Mamoulis, and D. Cheung. Mining frequent spatio-temporal sequential patterns. *Proc. of IEEE ICDM*, pages 82–89, 2005.

[2] M. Celik, S. Shekhar, J. Rogers, and J. Shine. Mixed-drove spatiotemporal co-occurrence pattern mining. *Proc. of IEEE TKDE*, 20(10):pages 1322–1335, 2008.

[3] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. *Proc. of ACM SIGKDD*, pages pages 330–339, 2007.

[4] J. Han, K. Koperski, and N. Stefanovic. Geominer: a system prototype for spatial data mining. In *Proc. of ACM SIGMOD*, SIGMOD '97, pages 553–556, 1997.

[5] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.-C. Hsu. Freespan: frequent pattern-projected sequential pattern mining. In *Proc. of ACM SIGKDD*, KDD '00, pages 355–359, 2000.

[6] Y. Huang, S. Shekhar, and H. Xiong. Discovering colocation patterns from spatial data sets: a general approach. *Proc. of IEEE TKDE*, 16(12):pages 1472–1485, Dec. 2004.

[7] Y. Huang, L. Zhang, and P. Zhang. A framework for mining sequential patterns from spatio-temporal event data sets. *Proc. of IEEE TKDE*, 20(4):pages 433–448, 2008.

[8] B. Mortazavi-Asl, H. Pinto, and U. Dayal. PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth. *Proc. of 17th International Conference on Data Engineering*, pages 215–224, 2000.

[9] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. Mining sequential patterns by pattern-growth: The prefixspan approach. *Proc. of IEEE TKDE*, 16(11):pages 1424–1440, 2004.

[10] S. Shekhar and Y. Huang. Discovering Spatial Co-Location Patterns A Summary Of Results. *Advances in Spatial and Temporal Databases*, pages 236–256, 2001.

[11] I. Tsoukatos and D. Gunopulos. Efficient mining of spatiotemporal patterns. *Advances in Spatial and Temporal Databases*, pages 425–442, 2001.

[12] J. Wang, W. Hsu, and M. Lee. Mining generalized spatio-temporal patterns. In *Database Systems for Advanced Applications*, pages 649–661. Springer, 2005.

[13] M. Yuan. In *Geographic Data Mining and Knowledge Discovery, Second Edition*, pages 347–365.