Vers une approche efficace d'extraction de motifs spatio-séquentiels

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

[*] Irstea, UMR TETIS, 500 rue Jean-François Breton, 34093 Montpellier - France
{prénom.nom}@teledetection.fr
[**] LIRMM, UMR 5506, 161 rue Ada, 34392 Montpellier - France
{prénom.nom}@lirmm.fr
[***] PPME - Université de la Nouvelle-Calédonie, BP R4, Nouméa, Nouvelle-Calédonie
{prénom.nom}@univ-nc.nc

Résumé. Ces dernières années, l'augmentation de la quantité d'informations spatio-temporelles stockées dans les bases de données a fait naître de nouveaux besoins, notamment en matière de gestion des risques naturels, sanitaires ou anthropiques (p. ex. compréhension de la dynamique d'une épidémie de Dengue). Dans cet article, nous définissons un cadre théorique pour l'extraction de motifs spatio-séquentiels, séquences de motifs spatiaux représentant l'évolution dans le temps d'une localisation et de son voisinage. Nous proposons un algorithme d'extraction efficace qui effectue un parcours en profondeur en s'appuyant sur des projections successives de la base de données. Nous introduisons également une mesure d'intérêt adaptée aux aspects spatio-temporels de ces motifs. Les expérimentations réalisées sur des jeux de données réels soulignent la pertinence de l'approche proposée par rapport aux méthodes de la littérature.

1 Introduction

Dans la vie quotidienne, nous pouvons observer de nombreux phénomènes se déroulant dans l'espace et dans le temps simultanément. Par exemple, les informations décrivant des trajets associent des informations spatiales (les coordonnées de la ville de départ et d'arrivée) et temporelles (l'heure de départ et d'arrivée). D'autres applications, avec des dynamiques plus complexes, sont beaucoup plus difficiles à analyser. Par exemple, dans le cas des épidémies de dengue, les experts en santé publique savent que l'évolution de la maladie dépend de facteurs environnementaux (p. ex. climat, proximité des zones infectées en points d'eau, mangroves...) et des interactions entre les humains et le vecteur de transmission (p. ex. le moustique qui transporte la maladie). Toutefois, l'impact des facteurs environnementaux et des interactions reste mal connu. Dans cet article, nous nous intéressons aux méthodes de découverte de connaissances permettant de mieux comprendre l'évolution des systèmes complexes pour la surveillance épidémiologique. Dans ce contexte, les méthodes de fouille de données spatio-temporelles apportent des solutions très pertinentes via l'identification sans hypothèse *a*

priori de relations entre variables et événements, caractérisées dans l'espace et dans le temps. Nous allons montrer dans l'état de l'art que les méthodes actuelles ne sont pas complètement adaptées à notre problématique. Elles ne permettent pas, notamment, d'étudier l'évolution de zones ou de quartiers en fonction de leur environnement propre et de celui de leurs voisins. Pour cette raison, nous avons défini un nouveau type de motifs que nous avons appelés motifs spatio-séquentiels, basés sur une extension des motifs séquentiels, de façon a considérer ensemble les dimensions spatial et temporel. Un exemple de motif dans le contexte de dengue est : *fréquemment, au cours des 10 dernières années, s'il pleut dans une zone et s'il y a de l'eau stagnante et des températures élevées dans des zones voisines, alors il y a une augmentation du nombre de moustiques dans des zones adjacentes, suivie par une augmentation du nombre de cas de dengue*. Ce motif peut être utilisé par les professionnels de la santé pour mieux comprendre comment les facteurs environnementaux influençent le développement des épidémies. Si ces motifs sont très intéressants d'un point de vue applicatif, ils sont aussi très difficiles à extraire en raison de la taille de l'espace de recherche associé. Proposer des méthodes passant à l'échelle constitue un véritable challenge. Pour cette raison, nous avons défini une mesure d'intérêt pour réduire le nombre de motifs générés et proposé un algorithme efficace basé sur des projections successives des données. Dans la section 2, nous listons des méthodes de fouille de données spatio-temporelles existantes et nous détaillons en quoi ces méthodes ne sont pas adaptées à notre problème. Dans la section 3, nous présentons le cadre théorique. Dans la section 4, nous introduisons notre algorithme appelé DFS-S2PMiner. Dans la section 5, nous présentons des expérimentations sur des données réelles. L'article finit par des conclusions et des perspectives.

2 Etat de l'art

Cet état de l'art se focalise sur les méthodes permettant d'analyser l'évolution et les interactions spatio-temporelles de caractéristiques d'objets ou d'évènements. Les travaux sur le suivi de trajectoires tels que Cao et al. (2005); Giannotti et al. (2007) ne seront pas étudiés car ils ne permettent pas de répondre à notre problématique applicative. Les premiers travaux ont traité les dimensions spatiales et temporelles séparément. Par exemple, Han et al. dans Han et al. (1997) et Shekhar et al. dans Shekhar et Huang (2001) ont recherché des *spatial patterns* ou *co-localisations*, c'est-à-dire des sous-ensembles de caractéristiques (object-type) fréquemment identifiés comme proches dans l'espace. Dans notre contexte, un exemple de co-localisation est : *dans un périmètre de 200 m, on trouve fréquemment des nids de moustiques près des étangs*. Dans un contexte autre que les bases de données spatiales, des auteurs comme Pei et al., dans Pei et al. (2004), se sont intéressés aux *sequential patterns* ou séquences temporelles, c'est-à-dire aux motifs variant dans le temps. Tsoukatos et al., dans Tsoukatos et Gunopulos (2001), ont étendu ces travaux pour représenter des ensembles de caractéristiques environnementales évoluant dans le temps. Ils extraient des séquences de caractéristiques qui apparaissent fréquemment dans des zones, mais sans prendre en compte le voisinage spatial. Un exemple de motif obtenu est : *dans de nombreuses régions, de fortes pluies surviennent avant la formation d'un étang, suivie par le développement de nids de moustiques (au sein de la même zone)*. Si ces approches, considérant uniquement les aspects spatiaux ou temporels sont sources d'informations pertinentes pour les experts, elles ne permettent pas de lier les caractéristiques de plusieurs zones dans le temps. Par exemple, elles ne permettent pas de

capter des relations du type : *Souvent, une forte pluie se produit avant la formation d'étangs suivie, dans une zone proche, du développement de nids de moustiques*. Dans Wang et al. (2005), Wang et al. se concentrent sur l'extraction de séquences représentant la propagation d'événements spatio-temporels dans des fenêtres temporelles prédéfinies. Ils introduisent deux concepts : les *Flow Patterns* et les *Generalized Spatiotemporal Patterns* afin d'extraire des séquences d'événements qui se produisent fréquemment à certains endroits. Ainsi, les auteurs vont pouvoir identifier des motifs de la forme : *des cas de dengue apparaissent fréquemment dans la région* $Z1$ *après la présence de températures élevées dans la région* $Z2$ *contenant des mares*. Ce type de motifs lient des zones précises comme $Z1$ et $Z2$ mais ils ne mettent pas en évidence de relation du type "près de", comme une "zone proche d'une autre". Huang et al., dans Huang et al. (2008), ont constaté que tous les motifs découverts avec les approches antérieures, ne sont pas pertinents car ils ne sont pas forcément *"denses"* dans l'espace et le temps. Ils ont donc proposé une mesure d'intérêt prenant en compte les aspects spatiaux et temporels pour extraire des séquences de caractéristiques globales. Cependant, ils étudient les événements d'une zone après l'autre. Ils ne prennent pas en compte les interactions telles que : *souvent, de fortes pluies et l'apparition d'étangs surviennent avant le développement de nids de moustiques*. Celik et al. Celik et al. (2006, 2008), ont proposé la notion de *Mixed-drove spatiotemporal co-occurrence patterns*, i.e. des sous-ensembles de deux ou plusieurs types d'événements dont les instances sont souvent proches dans le temps et l'espace, par exemple les événements-type : *cas de dengue, eau stagnante, haute température* apparaissent fréquemment dans des zones (instances) voisines et à des périodes différentes. Pour des raisons similaires à celles de Huang, Celik et al. ont proposé une mesure d'intérêt monotone, basée sur des mesures de prévalence spatiales et temporelles. Cependant, ils ne peuvent pas extraire les évolutions d'évènements-type au cours du temps (événements où chaque instance se produit nécessairement dans la même plage temporelle). Par exemple, ils n'extraient pas des motifs tels que : *forte pluie* dans une zone à un temps t_1, suivie d'*apparition d'étangs* dans la même zone au temps t_2 avec des cas de dengue dans une zone voisine induit l'apparition de *cas de dengue* dans la zone en question au temps t_3. Finalement, les approches proposées par Wang, Huang et Celik ne permettent pas de capturer l'évolution des zones en considérant l'ensemble des événement-types qui s'y déroulent ainsi que dans les zones voisines.

Dans cet article, nous décrivons une méthode d'extraction des séquences de motifs spatio-séquentiels (i.e. séquences d'ensembles d'événements spatiaux) appelées *S2P* (Spatio-Sequential Patterns). Nous avons pour objectif d'identifier des relations telles que : *la présence de cas de dengue dans une région est souvent précédée de températures élevées dans une zone située près de réservoirs d'eau*. Nous allons donc traiter les évolutions et les interactions entre la zone d'étude et son environnement immédiat. Par ailleurs, comme ce type de motifs est très difficile à extraire en raison du très grand espace de recherche généré, nous allons introduire une mesure d'intérêt pour assurer le passage à l'échelle de notre approche.

3 Motifs spatio-séquentiels : concepts et définitions

3.1 Définitions préliminaires

Une base de données spatio-temporelles est un ensemble structuré d'informations incluant des composantes géographiques (p. ex. des quartiers, des rivières...) et des compo-

santes temporelles (p. ex. pluie, vent...). Une telle base peut être définie comme un triplet $BD = \{D_S, D_T, D_A\}$ où D_T est la dimension temporelle, D_S la dimension spatiale et $\mathcal{D}_{\mathcal{A}}$= $\{D_{A_1}, D_{A_2}, \ldots D_{A_p}\}$ l'ensemble des dimensions qui décrivent les autres attributs.

La *dimension temporelle* est associée à un domaine de valeurs ordonnées dénoté $dom(D_T) = \{T_1, T_2, \ldots, T_t\}$ où T_i pour $i \in [1..t]$ est une *estampille temporelle* et $T_1 < T_2 < \ldots < T_t$.

La *dimension spatiale* est associée à un domaine de valeurs dénoté $dom(D_S) = \{Z_1, Z_2, \ldots Z_l\}$ où chaque Z_i pour $i \in [1..l]$ est une *zone*. Les zones sont liées par une relation de voisinage notée $voisin$ définie par : $voisin(Z_i, Z_j) = vrai$ si Z_i et Z_j sont voisines, $faux$ sinon.

Chaque *dimension d'analyse* D_{A_i} pour $i \in [1..p]$ est associée à un domaine de valeurs dénoté $dom(D_{A_i})$. Dans ces domaines, les valeurs peuvent être ordonnées ou non.

Pour illustrer les définitions, nous utilisons l'exemple du tableau 1 qui représente une base de données météo associée à trois villes sur trois jours consécutifs. Le tableau contient la température, les précipitations, la force du vent et la vitesse des rafales en Km/h. Les trois villes sont liées par une relation de proximité décrite dans la figure 1.

Ville	Date	Température	Précipitation	Vent	Rafales
Z_1	22/12/10	T_m	P_m	V_m	-
Z_1	23/12/10	T_m	P_m	V_l	-
Z_1	24/12/10	T_l	P_m	V_m	55
Z_2	22/12/10	T_m	P_m	V_m	-
Z_2	23/12/10	T_l	P_m	V_l	-
Z_2	24/12/10	T_l	P_l	V_m	-
Z_3	22/12/10	T_l	P_m	V_s	75
Z_3	23/12/10	T_m	P_s	V_l	-
Z_3	24/12/10	T_l	P_s	V_s	55
...	...	...	...	...	...

Tab. 1: Changements climatiques dans trois zones : Z_1, Z_2 et Z_3 pour le 22, 23, 24 décembre 2010.

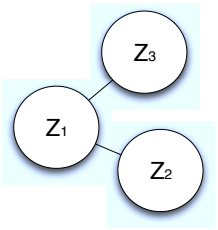


Fig. 1: Villes voisines.

Dans le tableau 1, $D_T = \{Date\}$, $D_S = \{Ville\}$ et D_A=*{Température, Précipitation, Vent, Rafales}*. Le domaine de la dimension temporelle est $dom(D_T) = \{22/12/10, 23/12/10, 24/12/10\}$ avec $22/12/10 < 23/12/10 < 24/12/10$. Le domaine de la dimension spatiale est $dom(D_S) = \{Z_1, Z_2, Z_3\}$ avec $voisin(Z_1, Z_2) = vrai$, $voisin(Z_1, Z_3) = vrai$ et $voisin(Z_2, Z_3) = faux$. Finalement, le domaine de la dimension d'analyse *Température* est *dom(Température)*= $\{T_l, T_m, T_s\}$ et de la dimension d'analyse *Rafales* est $dom(Rafales) = \{55, 75\}$.

3.2 Motifs spatio séquentiels

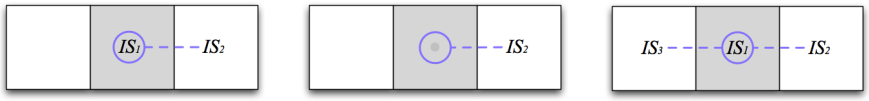
Définition 1 Item et Itemset. Soit I un *item*, une valeur littérale associée à une dimension D_{A_i}, $I \in dom(D_{A_i})$. Un *itemset* $IS = (I_1 I_2 \ldots I_n)$ avec $n \leq p$, est un ensemble non vide d'*items* tel que $\forall i, j \in [1..n], \exists k, k' \in [1..p]$, $I_i \in dom(D_{A_k})$, $I_j \in dom(A_{D_{A_{k'}}})$ et $k \neq k'$.

Nous définissons la relation *In* entre zones et itemsets qui décrit l'apparition de l'itemset *IS* dans la zone Z à l'instant t dans la base de données BD : $In(IS, Z, T)$ est vraie si IS est présent dans la zone Z au temps t dans BD. Par exemple, soit l'itemset $IS = (T_m P_m V_l)$ alors,

$In(IS, Z_1, 23/12/2010) = vrai$ représente l'apparition de l'itemset $(T_m P_m V_l)$ dans la zone Z_1 à la date 23/12/2010 (voir tableau 1). Nous définissons maintenant la notion d'*interaction* entre zones voisines.

Définition 2 Itemset spatial. Soient deux itemsets IS_i, IS_j, il existe une *proximité spatiale* entre IS_i et IS_j si et seulement si $\exists Z_i, Z_j \in dom(D_S)$, $\exists t \in dom(D_T)$ tels que $In(IS_i, Z_i, t) \wedge In(IS_j, Z_j, t) \wedge voisin(Z_i, Z_j)$ est $vrai$. Deux itemsets IS_i et IS_j qui sont proches spatialement, forment un *itemset spatial* noté $I_{ST} = IS_i \cdot IS_j$.

Afin de faciliter les notations, nous introduisons un *opérateur de groupement* $n\text{-}aire$ noté *[ ]*, qui permet de regrouper une liste d'itemsets affectés par l'opérateur $\cdot$ (*à côté de*). Nous utilisons également le symbole θ pour représenter l'*absence d'itemsets* dans une zone. La figure 2 montre les 3 types d'itemsets spatiaux que nous pouvons construire avec les notations introduites précédemment. Les lignes pointillées représentent la dynamique spatiale.



(a) (b) (c)

Fig. 2: Représentation graphique des itemsets spatiaux (a) $IS_1 \cdot IS_2$ (b) $\theta \cdot IS_2$ (c) $IS_1 \cdot [IS_2; IS_3]$.

L'itemset spatial $I_{ST} = (T_m \cdot [V_l; P_m])$ signifie qu'il existe une zone pour laquelle l'événement T_m s'est produit et que l'évènement V_l et l'événement P_m se sont produits au même temps dans des zones voisines. L'itemset spatial $I_{ST} = (\theta \cdot [T_m; P_l])$ indique qu'il existe une zone où il n'y a eu aucun événement mais dans deux zones voisines distinctes se sont produits respectivement les événements T_m et P_l au même moment.

Définition 3 Inclusion d'itemset spatial. Un itemset spatial $I_{ST} = IS_i \cdot IS_j$ est *inclus*, avec l'operateur noté $\subseteq$, dans un autre itemset spatial $I'_{ST} = IS'_k \cdot IS'_l$, si et seulement si $IS_i \subseteq IS'_k$ et $IS_j \subseteq IS'_l$.

Par exemple, l'itemset spatial $I_{ST} = (T_m P_m \cdot V_l)$ est *inclus* dans l'itemset spatial $I'_{ST} = (T_m P_m \cdot V_l 55)$ car $(T_m P_m) \subseteq (T_m P_m)$ et $(V_l) \subseteq (V_l 55)$.

Définition 4 Séquence spatiale. Une *séquence spatiale* ou ***S2*** est une liste ordonnée d'itemsets spatiaux, notée $s = \langle I_{ST_1} I_{ST_2} \ldots I_{ST_m} \rangle$ où $I_{ST_i}, I_{ST_{i+1}}$ respectent la contrainte de séquentialité temporelle pour tout $i \in [1..m-1]$.

Un exemple de séquence spatiale est $s = \langle (T_m)(\theta \cdot [P_l; V_s])(V_l \cdot [P_l; T_l]) \rangle$ qui est illustré en figure 3. Les flèches représentent la dynamique temporelle et les lignes pointillés représentent la relation de voisinage spatial entre itemsets.

Une relation de généralisation (ou spécialisation) entre séquences spatiales est définie de la manière suivante :

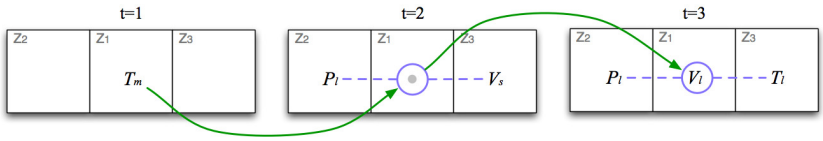


Fig. 3: Exemple de dynamique spatio-temporelle.

Définition 5 Inclusion de *S2*. Une séquence spatiale $s = \langle I_{ST_1} I_{ST_2} \dots I_{ST_m} \rangle$ est plus spécifique qu'une séquence spatiale $s' = \langle I'_{ST_1} I'_{ST_2} \dots I'_{ST_n} \rangle$, notée $s \preceq s'$, s'il existe des entiers $j_1 \leq \dots \leq j_m$ tels que $I_{ST_1} \subseteq I'_{ST_{j_1}}, I_{ST_2} \subseteq I'_{ST_{j_2}}, \dots, I_{ST_m} \subseteq I'_{ST_{j_m}}$.

La séquence spatiale $s = \langle (T_l P_m \cdot P_l V_s)(55) \rangle$ est incluse dans la séquence spatiale s' = $\langle (T_l P_m \cdot P_l V_s)(55 \cdot V_s) \rangle$ car $(T_l P_m \cdot P_l V_s) \subseteq (T_l P_m \cdot P_l V_s)$ et $(55) \subseteq (55 \cdot V_s)$.

Définition 6 Support d'une séquence spatiale. Pour une zone spécifique Z, on note s_Z la séquence de données spatiales associée dans la base de données BD. s_Z contient ou supporte une séquence spatiale s si s est une sous séquence de s_Z. Le support d'une séquence spatiale s, noté par *supp(s)*, est ainsi défini comme le nombre de zones qui supportent s, i.e. $supp(s) = | \{Z \in dom(D_S) \mid s \subseteq s_Z\} |$. Si le support de la séquence spatiale est supérieur à un seuil défini par l'utilisateur, la séquence est fréquente et correspond à un *motif spatio-séquentiel **(S2P)***.

Néanmoins, dans un contexte spatio-temporel, nous avons besoin de définir une mesure de prévalence plus précise et adaptée, comme nous l'expliquerons dans la section suivante.

3.3 Indice de participation spatio-temporel

La notion de motif spatio-séquentiel proposée permet de prendre en compte les aspects spatial et temporel. Pour gérer de manière efficace l'exploitation de tels motifs, nous avons défini une nouvelle mesure de filtrage. Afin de souligner la participation d'un élément dans une séquence spatiale, nous proposons une adaptation de l'indice de participation défini dans Huang et al. (2004) comme la combinaison de deux mesures : *l'indice de participation spatiale* et *l'indice de participation temporelle*, tenant compte respectivement de la répartition spatiale et du nombre d'occurrences dans le temps.

Définition 7 Ratio de participation spatial. Soit la séquence spatiale s et $I \in dom(D_A)$ un item de s, le *ratio de participation spatial* de s par rapport à I noté $SPr(s, I)$ est le nombre de zones contenant s sur le nombre total de zones où l'item I apparaît dans la base.

$$SPr(s, I) = \frac{supp(s)}{supp(I)}$$

Définition 8 Indice de participation spatial. Soit la séquence spatiale $s = \langle I_{ST_1}, I_{ST_2}, \dots I_{ST_n} \rangle$, l'*indice de participation spatial* de s noté $SPi(s)$ est le minimum de ses *ratio de participation spatial* :

$$SPi(s) = MIN_{\forall I \in Dom(A), I \in s} \{SPr(s, I)\}$$

Définition 9 Ratio de participation temporel. Soit la séquence spatiale s et $I \in Dom(A)$ un item de s, le *ratio de participation temporel* de s par rapport à I noté $TPr(s, I)$ est le nombre d'occurrences de s sur le nombre total d'occurrences de I dans la base.

$$TPr(s, I) = \frac{NbOccurrences(s)}{NbOccurrences(I)}$$

Définition 10 Indice de participation temporel. Soit la séquence spatiale $s = \langle I_{ST_1}, I_{ST_2}, \ldots I_{ST_n} \rangle$, l'*indice de participation temporel* de s noté $TPi(s)$ est le minimum de ses *ratio de participation temporel* :

$$TPi(s) = MIN_{\forall I \in Dom(A_i), I \in s} \{TPr(s, I)\}$$

Nous définissons l'*indice de participation spatio-temporel* d'une séquence spatiale s, noté $STPi(s)$, comme le produit pondéré des deux mesures précédentes :

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

La problématique d'extraction de motifs spatio-séquentiels à partir d'une base de données spatio-temporelles BD consiste à retrouver toutes les séquences spatiales dont l'indice de participation spatio-temporel est supérieur à un seuil spécifié par l'utilisateur *supp_min*. Notons que le prédicat "STP_i est supérieur à un seuil de l'utilisateur" est anti-monone. Si un motif spatio-séquentiel s n'est pas fréquent, alors tous les motifs s' tels que $s \preceq s'$, sont également non fréquents. Cette propriété est utilisée dans notre algorithme d'extraction de motifs pour élaguer l'espace de recherche et trouver rapidement les motifs spatio-séquentiels fréquents.

4 Extraction des motifs spatio-séquentiels

Dans cette section, nous proposons un algorithme appelé DFS-S2PMiner pour extraire des motifs spatio-séquentiels considérant à la fois les aspects spatiaux et temporels. DFS-S2PMiner adopte une stratégie du type *depth-first-search* basée sur des projections successives de la base de données (à des fins d'évolutivité) comme le font FP-Growth Han et al. (2000) et Prefixspan Mortazavi-Asl et al. (2000). Plus précisément, cet algorithme est basé sur la stratégie *pattern growth* utilisée dans Han et al. (2000). Le principe de cette approche est d'extraire des motifs fréquents sans l'étape de génération des candidats. Cette approche fait récursivement des projections de la base de données, les associe à un fragment de motif fréquent et "fouille" chacune des bases projetées séparément. Les motifs fréquents sont ainsi étendus progressivement suivant un parcours en profondeur de l'espace de recherche. Notre approche est décrite par l'algorithme 1 et fait appel à la procédure récursive *Prefix-growthST* (algorithme 2). Cette procédure commence par rechercher l'ensemble F_1 des items fréquents I et $\theta \cdot I$ de la projection $BD|_\alpha$ (ligne 1 de l'algorithme 2). Ces items vont constituer les extensions possibles de la séquence α. Notons que lors du premier appel récursif, nous avons $\alpha = \{\}$. Dans ce cas, la projection $BD|_\alpha$ correspond à la base de données spatio-temporelles BD de départ. Ensuite, pour chaque item ou item spatial fréquent $X \in F_1$, nous étendons le motif spatio-séquentiel α avec X (lignes 3-4). Pour faire cela, nous avons deux possibilités : 1) ajouter X au dernier itemset spatial de la séquence α (ligne 3) ou 2) insérer X après (i.e. au temps suivant) le

dernier itemset spatial de α (ligne 4). Une fois l'extension du motif effectuée, nous vérifions le support des séquences spatiales résultantes (lignes 5 et 9). Chaque motif spatio-séquentiel (séquence spatiale fréquente) trouvé est enregistré dans l'ensemble des solutions F (lignes 6 et 10). L'algorithme effectue ensuite une projection de la base de données par rapport à ces motifs. Les appels récursifs qui suivent permettent de construire les sur-séquences des motifs fréquents trouvés à partir des bases projetées correspondantes (lignes 7 et 11). L'algorithme s'arrête lorsqu'il n'y a plus de projections qui puissent être générées.

Algorithme 1 DFS-S2PMiner

ENTRÉE : Une base de données spatio-temporelles BD et un support minimal $supp\_min$
SORTIE : L'ensemble de motifs spatio-séquentiels fréquents F
1: $\alpha \leftarrow \{\}$
2: Appeller *Prefix-growthST*$(\alpha, supp\_min, BD|_\alpha, F)$

Algorithme 2 Prefix-growthST$(\alpha, supp_min, BD|_\alpha, F)$

ENTRÉE : une séquence spatiale α , le support minimal $supp\_min$, la projection $BD|_\alpha$ de la base de données spatio-temporelles par rapport à α, et F un ensemble de motifs spatio-séquentiels fréquents ;
1: $F_1 \leftarrow$ {l'ensemble d'items I et $\theta.I$ fréquents dans $BD|_\alpha$, avec $I \in \bigcup_{i\in[1..p]} dom(D_{A_i})$ }
2: **pour tout** $X \in F_1$ **faire**
3: $\beta \leftarrow \alpha X$
4: $\delta \leftarrow \alpha(X)$
5: **si** $STPi(\beta) \geq supp\_min$ **alors**
6: $F \leftarrow F \cup \beta$;
7: Prefix-growthST$(\beta, supp_min, BD|_\beta, F)$
8: **finsi**
9: **si** $STPi(\delta) \geq supp\_min$ **alors**
10: $F \leftarrow F \cup \delta$;
11: Prefix-growthST$(\delta, supp_min, BD|_\delta, F)$
12: **finsi**
13: **fin pour**

Exemple : Nous illustrons cet algorithme en utilisant des données du tableau 1 et la figure 1 avec un support minimal de $2/3$. Dans un premier temps, *Prefix-growthST* (cf. Algorithme 2) commence par extraire les items fréquents et items spatiaux fréquents de BD (ligne 1), soit :

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

Pour chaque item fréquent I et $\theta \cdot I$ trouvé (ligne 2), l'algorithme calcule la projection de la base de données par rapport à ces items (aucune extension n'est faite ici car α est vide). Par exemple, pour l'item fréquent P_m, nous obtenons la projection suivante :

Zones	Séquences	Zones voisines	Séquences voisines
Z_1	$S_1 = \langle(\_\boldsymbol{V_m})(T_m\boldsymbol{P_m}\boldsymbol{V_m}55)\rangle$	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	$S_1 = \langle(\_V_m)(T_mP_mV_l)(T_lP_mV_m55)\rangle$

Tab. 2: Projection de $\langle(P_m)\rangle$ sur BD.

Chacune de ces bases projetées est ensuite utilisée dans un appel récursif permettant de rechercher des sur-séquences fréquentes (lignes 7 et 11). Le premier appel récursif va construire

les sous-séquences ayant pour préfixe $\langle(P_m)\rangle$ à partir de la base projetée décrite dans le tableau 2. Plus précisément, l'algorithme va d'abord rechercher les items fréquents dans cette base projetée (lignes 1), puis étendre $\langle(P_m)\rangle$. Les items fréquents obtenus pour $BD|_{\langle(P_m)\rangle}$ sont : $\{V_m : 2, T_m : 2, P_m : 2, V_l : 3, T_l : 3, 55 : 2, \theta \cdot V_m : 3, \theta \cdot T_l : 3, \theta \cdot P_m : 3, \theta \cdot V_l : 3, \theta \cdot T_m : 3, \theta \cdot 55 : 3\}$. Le premier item fréquent trouvé est $\langle V_m\rangle : 2$. On obtient donc les motifs spatio-séquentiels $\langle(P_m V_m)\rangle$ (ligne 3) et $\langle(P_m)(V_m)\rangle$ (ligne 4). Seul $\langle(P_m)(V_m)\rangle$ avec $STPi = 2/3$ est fréquent (ligne 9), l'algorithme utilise ce motif pour faire une nouvelle projection (cf. tableau 3) et rechercher récursivement toutes les sous-séquences fréquentes ayant pour préfixe $\langle(P_m)(V_m)\rangle$.

Zones	Séquences	Zones voisines	Séquences voisines
Z_1	$S_1 = \langle(T_m P_m V_l)(T_l P_m V_m 55)\rangle$	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	$S_2 = \langle(T_l P_m V_l)(T_l P_l V_m)\rangle$	Z_1	$S_1 = \langle(T_m \boldsymbol{P_m} V_l)(T_l \boldsymbol{P_m} V_m 55)\rangle$
Z_3	$S_3 = \emptyset$	Z_1	$S_1 = \langle(T_m \boldsymbol{P_m} V_l)(T_l \boldsymbol{P_m} V_m 55)\rangle$

Tab. 3: Projection de $\langle(P_m)(V_m)\rangle$ sur BD.

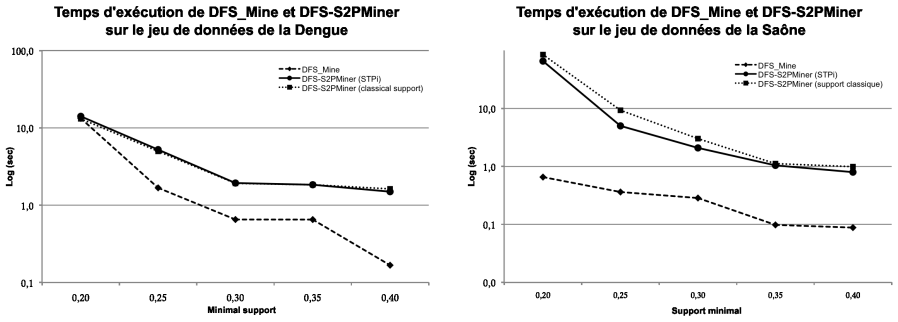
L'item spatial $\theta \cdot P_m : 2$ est un des items fréquents dans l'appel récursif qui suit, donc l'algorithme construit le motif spatio-séquentiel fréquent $\langle(P_m)(V_m)(\theta \cdot P_m)\rangle$ avec un $STPi = 1$. Une fois tous les items fréquents projetés, l'algorithme va parcourir une autre "branche" de l'espace de recherche, i.e. les motifs commençant par $\langle(T_m)\rangle$ (voir l'ensemble F_1). L'algorithme procède donc globalement de la même manière que les items soient spatiaux ou non. La principale différence se situe dans la manière de compter le support. En effet, le support d'un item spatial représente le nombre de zones ayant au moins une fois dans leur voisinage l'item en question (c'est pour cette raison que nous avons $\theta \cdot V_l : 3$ dans le tableau 2). Notons que lorsque l'algorithme étend un motif du type $\langle(I_{ST_1})(I_{ST_2})...(IS_k \cdot X)\rangle$ avec un item fréquent $\theta \cdot Y$, l'opérateur de groupement n-aire est utilisé afin de représenter la séquence sous la forme $\langle(I_{ST_1})(I_{ST_2})...(IS_k \cdot [X;Y])\rangle$.

5 Résultats expérimentaux

DFS-S2PMiner a été développé en Java et testé sur deux jeux de données réelles. La première base de données spatio-temporelles concerne des données épidémiologiques de suivi de la dengue. Ces données ont été collectées à *Nouméa (Nouvelle Calédonie)* [1] sur un territoire divisée en 81 quartiers couvrant $45,7\ km^2$ et correspondent à 26 dates pour lesquelles nous disposons d'informations discrétisées décrivant les caractéristiques associées à chaque quartier. La deuxième base de données est constituée de relevés d'indicateurs biologiques dans des rivières de la Saône, comme par exemple, l'IBGN (Indice Biologique Global Normalisé) et l'IBD (Indice Biologique Diatomée). Ces données hydrologiques sont associées, d'abord aux stations de relevés positionnées sur les cours d'eau puis aux relevés effectués par les stations stratégiquement positionnées le long des bassins versants. Ce jeu correspond à 223 zones, 815 dates et 10 attributs.

1. Données issues de la base de données de la Direction des Affaires Sanitaires et Sociale de la Nouvelle Calédonie, de l'Institut Pasteur, de l'IRD et de l'UNC (Convention 2010)

Nous avons choisi de comparer notre approche avec la méthode la plus proche sémantiquement : l'algorithme DFS_Mine proposé par Tsoukatos Tsoukatos et Gunopulos (2001). En effet, cet algorithme extrait des séquences d'itemsets représentant l'évolution d'une zone mais sans prendre en compte les zones voisines. Des expérimentations ont été effectuées avec un processeur Intel Core i5 avec 4G de RAM sur Linux.

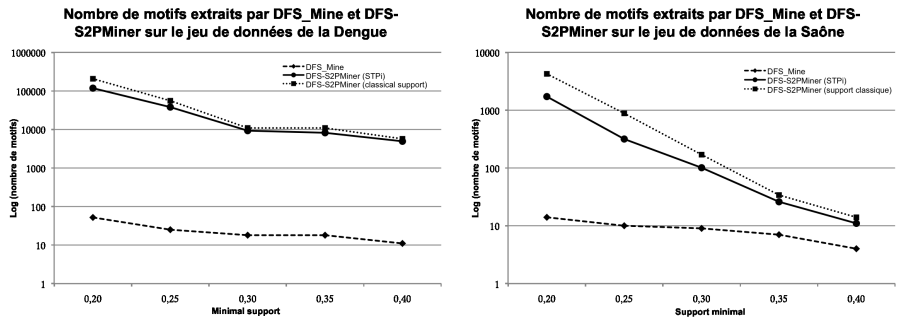


(a) (b)

Fig. 4: Temps d'exécution de DFS_Mine et DFS-S2PMiner sur les données (a) de la Dengue (b) des rivières.

D'un point de vue qualitative des résultats, nous avons comparé les motifs obtenus par notre approche avec ceux obtenus par DFS_Mine sur l'ensemble des données de la Dengue. Par exemple, les deux approches trouvent des motifs séquentiels classiques tels que *peu de piscines, peu de précipitations et faible présence de cimetières suivi par quelques cas de dengue, peu de précipitations et du vent*. Cependant, notre approche trouve également des motifs complexes tels que *peu de piscines, peu de précipitations et faible présence de cimetières suivi par peu de piscines et beaucoup de précipitations dans deux zones voisines, suivi par présence de dengue dans une zone voisine à la zone d'étude*. Cet exemple donne une idée de la richesse des motifs extraits par notre approche en mettant en évidence l'influence des zones voisines. Lorsque l'indice de participation spatio-temporel (STPi) est utilisé en tant que mesure d'intérêt, nous ne pouvons pas faire des comparaisons avec d'autres approches, car les mesures de prévalence sont différentes. L'approche de Tsoukatos propose des séquences se produisant dans de nombreuses zones, mais pas nécessairement à plusieurs dates. Notre approche, quant à elle, extrait des motifs se produisant dans de nombreuses zones et à plusieurs dates. L'intérêt de notre proposition est de considérer le poids temporel de motifs extraits.

D'autre part, une évaluation quantitative de notre approche a été faite. Nous avons comparé les temps d'exécution de notre algorithme à celui de DFS_Mine. La figure 4 montre les temps d'exécution des deux algorithmes DFS_Mine (support classique) et DFS-S2PMiner (en utilisant le support classique et l'indice de participation spatio-temporel) sur les deux jeux de données et pour plusieurs seuils. Les temps d'exécution sont relativement similaires, tandis que notre approche fait un traitement plus complexe. En effet, comme le montre la figure 5, la mesure proposée, STPi, permet un élagage efficace de l'espace de recherche, même pour le grand jeu de données des rivières.



(a) (b)

Fig. 5: Nombre de motifs S2P extraits par DFS_Mine et DFS-S2PMiner sur les données (a) de la dengue (b) des bassins versants de la Saône.

6 Conclusion et perspectives

Dans cet article, nous avons proposé un nouveau concept de motifs spatio-temporels appelés motifs spatio-sequentiels ***(S2P)***. Ces motifs permettent d'étudier l'évolution, d'un ensemble de caractéristiques "spatialisées", dans le temps en prenant en compte l'environnement voisin. Cette évolution décrit bien, par exemple, l'évolution de la dengue dans le temps en fonction des caractéristiques des quartiers et leur voisinage. Un cadre formel a été établi pour définir les motifs S2P de manière générique. Pour extraire ces motifs, nous avons proposé un algorithme appelé *DFS-S2PMiner* qui suit une stratégie de recherche en profondeur *(depth-first search)*. Une nouvelle mesure de prévalence a été définie pour faire face aux limites du support classique en considérant les aspects spatiaux et temporels. Nous avons testé notre méthode sur deux jeux de données réelles. Les résultats montrent l'intérêt de l'approche pour extraire efficacement des motifs spatio-temporels très riches. Les perspectives associées à ce travail sont nombreuses. Tout d'abord, nous souhaitons étendre la notion de voisinage à des voisins situés à une distance n de la zone concernée tout en permettant le passage à l'échelle. Il ne sera pas nécessaire de redéfinir les concepts mais de proposer une heuristique efficace de parcours de l'espace de recherche.

Références

Cao, H., N. Mamoulis, et D. Cheung (2005). Mining frequent spatio-temporal sequential patterns. *Fifth IEEE International Conference on Data Mining (ICDM'05)* (ii), 82–89.

Celik, M., S. Shekhar, J. Rogers, et J. Shine (2006). Sustained Emerging Spatio-Temporal Co-occurrence Pattern Mining : A Summary of Results. *18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)*, 106–115.

Celik, M., S. Shekhar, J. Rogers, et J. Shine (2008). Mixed-drove spatiotemporal co-occurrence pattern mining. *IEEE Transactions on Knowledge and Data Engineering 20*(10), 1322–1335.

Giannotti, F., M. Nanni, F. Pinelli, et D. Pedreschi (2007). Trajectory pattern mining. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD'07*, 330.

Han, J., K. Koperski, et N. Stefanovic (1997). Geominer : a system prototype for spatial data mining. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, SIGMOD'97, NY, USA, pp. 553–556. ACM.

Han, J., J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, et M.-C. Hsu (2000). Freespan : frequent pattern-projected sequential pattern mining. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD'00, NY, USA, pp. 355–359. ACM.

Huang, Y., S. Shekhar, et H. Xiong (2004). Discovering colocation patterns from spatial data sets : a general approach. *IEEE Transactions on Knowledge and Data Engineering 16*(12), 1472–1485.

Huang, Y., L. Zhang, et P. Zhang (2008). A framework for mining sequential patterns from spatio-temporal event data sets. *IEEE Transactions on Knowledge and Data Engineering 20*(4), 433–448.

Mortazavi-Asl, B., H. Pinto, et U. Dayal (2000). PrefixSpan, : mining sequential patterns efficiently by prefix-projected pattern growth. *Proceedings 17th International Conference on Data Engineering*, 215–224.

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

Shekhar, S. et Y. Huang (2001). Discovering Spatial Co-Location Patterns A Summary Of Results. *Advances in Spatial and Temporal Databases*, 236–256.

Tsoukatos, I. et D. Gunopulos (2001). Efficient mining of spatiotemporal patterns. *Advances in Spatial and Temporal Databases*, 425–442.

Wang, J., W. Hsu, et M. Lee (2005). Mining generalized spatio-temporal patterns. In *Database Systems for Advanced Applications*, pp. 649–661. Springer.

Summary

In these last years, large quantity of spatio-temporal data stored leads to new needs such that management of natural risks, health or anthropogenic (e.g. understanding the dynamic of dengue epidemic). In this paper, we define a new theoretical framework for extracting spatio-sequential patterns. A spatio-sequential pattern is a sequence representing evolution of locations and their neighborhoods over time. We propose an efficient algorithm based on depth-first-search with successive projections over the database. We introduce a new interestingness measure taking into account both spatial and temporal aspects. Experiments are conducted on real datasets highlight the relevance of our method.