Extraction de motifs spatio-temporels à différentes échelles avec gestion de relations spatiales qualitatives

Mickaël Fabrègue[*,**], Agnès Braud[**], Sandra Bringay[***,*****], Florence Le Ber[****,******], Maguelonne Teisseire[*,***]**

[*] *TETIS, Irstea, 500 Rue Jean-François Breton, 34000 Montpellier*
[**] *LSIIT, CNRS-Uds, Pôle API Bd Sébastien Brant, 67412 Illkirch*
[***] *LIRMM UM2 CNRS, UMR 5506 - CC 477, 161 rue Ada, 34095 Montpellier*
[****] *LHYGES ; Université de Strasbourg/ENGEES, CNRS ; 67000 Strasbourg*
[*****] *MIAp UM3, Université Paul-Valery, Route de Mende, 34199 Montpellier*
[******] *LORIA UMR 7503, 54500 Vandœuvre-lès-Nancy*

RÉSUMÉ. *Les bases de données géoréférencées contiennent un important volume de données temporelles et spatiales et sont particulièrement utilisées dans le cadre d'analyses environnementales. Plusieurs méthodes ont été proposées pour l'exploration de telles bases de données, mais aucune ne permet d'exploiter toute la richesse des données, en particulier leurs dimensions spatiales et temporelles. Dans cet article, nous introduisons un nouveau type de motifs spatio-temporels considérant les relations entre objets spatiaux mais aussi les différentes échelles géographiques. Nous proposons un algorithme d'extraction de motifs STR_PrefixGrowth applicable sur un important volume de données. Nous traitons un exemple de données hydrobiologiques collectées sur le bassin de la Saône durant les 19 dernières années. Les expérimentations menées soulignent l'intérêt de notre méthode par rapport aux méthodes existantes.*

ABSTRACT. *Georeferenced databases contain a huge volume of temporal and spatial data. They are notably used in environmental analysis. Several works address the problem of mining those data, but none are able to take into account the richness of the data and especially their spatial and temporal dimensions. In this paper, we focus on the extraction of a new kind of spatio-temporal patterns which consider the relationship between spatial objects and also various geographical scales. We propose an algorithm, STR_PrefixGrowth, which can be applied on a huge amont of data. The proposed method is evaluated on hydrological data collected on the Saône basin during the last 19 years. Our experiments emphasize the contribution of our approach toward the existing methods.*

MOTS-CLÉS : *Fouille de données, Motifs séquentiels, Spatio-temporel, Écosystème aquatique*

KEYWORDS: *Data Mining, Sequential patterns, Spatio-temporal, Aquatic ecosystem*

1. Introduction

L'explosion récente des technologies mobiles et des données géoréférencées a fait émerger un nouveau type de données, qualifiées de spatio-temporelles. À chaque donnée est associée une information spatiale (i.e. une localisation) et une information temporelle (i.e. une date). De nouveaux besoins ont fait leur apparition comme le suivi d'événements dans le temps et l'espace. Les domaines concernés sont par exemple la propagation de l'information dans les réseaux sociaux [Lin *et al.*2011], la surveillance des épidémies [Gubler2002] ou comme dans cet article, le suivi hydrologique des cours d'eau. Dans ces domaines, le volume des données est très souvent important et les informations hétérogènes. L'aspect géographique des données peut être associé à une sémantique fine basée sur l'inclusion mais également sur d'autres relations variant selon les domaines d'application. En effet, les objets géographiques sont souvent décrits selon différents niveaux de granularité spatiale. Une zone peut être incluse dans une autre zone (cf. la région Languedoc Roussillon est divisée en départements Aude, Gard, Hérault...). De plus, les objets géographiques sont liés par des relations spatiales. Par exemple, une zone peut être à côté d'une autre zone ou orientée au nord ou à l'est d'une autre zone (cf. le Gard et l'Hérault sont deux zones voisines et le Gard est au nord de l'Hérault). Dans cet article, nous allons nous intéresser aux méthodes de fouille de données qui prennent en compte la temporalité mais aussi la richesse de la sémantique que l'on peut associer aux liens existant entre les objets géographiques tels que nous venons de les décrire. Notre objectif est de proposer une méthode pour extraire des motifs spatio-temporels mettant en évidence des comportements fréquents et qui soit applicable sur de gros volumes de données. C'est dans le contexte de l'environnement et plus particulièrement celui des écosystèmes aquatiques qu'a été appliquée la méthode proposée. Le jeu de données réelles est issu de prélèvements hydrologiques sur le bassin versant de la Saône, utilisé dans le cadre de l'ANR Fresqueau. Ce projet vise à proposer des outils opérationnels pour étudier l'état des systèmes aquatiques dans le cadre de la Directive Cadre Européenne sur l'eau (DCE, 2000), avec pour objectif en 2015 d'assurer le bon état des eaux des milieux aquatiques et des bassins versants.

2. État de l'art

L'extraction de motifs a fait l'objet de nombreuses recherches dans le domaine de la fouille de données. La découverte de motifs repose sur la mise en évidence d'une information récurrente dans des données, caractérisant un comportement fréquent. Cette connaissance peut prendre différentes formes.

Plusieurs auteurs proposent des motifs qui prennent en compte le temps et l'espace. Dans [Wang *et al.*2005], les données sont représentées sous la forme d'un ensemble de grilles dans lesquelles apparaissent des évènements (des items), chaque grille représentant l'état de la grille spatiale à un instant t. Pour chaque date et chaque position absolue, un itemset, ensemble d'évènements (d'items) est généré. Pour chaque position, une séquence d'itemsets est alors construite en considérant toutes les dates. Les

motifs séquentiels sont alors extraits à partir de cet ensemble de séquences, en utilisant une position absolue comme point de référence. Un exemple de motif obtenu est $\langle$(Pluie(0,0))(Humidité(0,1))$\rangle$, qui signifie que l'on trouve fréquemment de la pluie aux coordonnées 0,0 et ensuite de l'humidité aux coordonnées 0,1. Ce type de motif a le désavantage d'être sensible au choix du point de référence et l'espace est limité à une représentation sous forme de grille.

Dans [Huang *et al.*2008], les auteurs proposent la notion d'événements proches dans le temps et l'espace. Une fenêtre temporelle et spatiale est définie par un intervalle de temps et par un intervalle de distance. Les motifs sont sous forme de règles d'associations du type $\langle$Pluie$\Rightarrow$Humidité$\rangle$ qui signifie que dans des zones proches et à des dates proches, on trouve de la pluie suivie par de l'humidité. Ces motifs n'expriment pas le fait que les objets spatiaux soient liés par une relation, ni la présence de plusieurs échelles géographiques.

Une extraction de motifs spatio-temporels en utilisant une relation de voisinage ou de proximité entre séquences est proposée dans [Alatrista-Salas *et al.*2012]. Les motifs sont de la forme $\langle$(Humidité .[Pluie Vent])(Humidité Pluie)$\rangle$. La relation de voisinage est mise en évidence avec l'opérateur de voisinage **.** et l'opérateur de groupement **[]**. Prenons l'exemple d'une ville où nous trouvons ce motif. Cela signifie qu'il y a eu de l'humidité à une certaine date et au même moment de la pluie et du vent dans une ville proche (par une distance euclidienne ou défini par l'utilisateur). Plus tard, il y a eu de l'humidité et de la pluie dans la ville. Cette relation spatiale reste simple et il n'est pas possible de la spécialiser (i.e. une relation représentée par un domaine de valeurs), ni d'avoir plusieurs granularités (i.e. plusieurs échelles géographiques). Elle se limite à un seul type de relation : la proximité dans l'espace.

Une gestion de la granularité sur l'espace est proposée dans [Tsoukatos et Gunopulos2001]. Comme dans [Wang *et al.*2005], la spatialité est représentée par une grille d'événements et la temporalité par un ensemble de grilles spatiales. L'utilisateur choisi un niveau de granularité qui va fusionner un ensemble de cases de la grille. Plus la valeur de granularité est élevée, plus les cases seront fusionnées, ce qui permet la généralisation des données d'un point de vue spatial. Pour extraire les motifs, il est donc nécessaire de donner une valeur de granularité et la spatialité se limite à une grille. De plus l'extraction revient à extraire des motifs séquentiels classiques du type $\langle$(Soleil)(Vent)(Soleil,Humidité=Faible)$\rangle$ qui se lit de la manière suivante : fréquemment l'événement Soleil est suivi de l'événement Vent, suivi des événements Soleil et Humidité=Faible pour un niveau de granularité précis.

Toutes ces méthodes ne permettent pas de traiter efficacement des données géographiques complexes, avec des objets géographiques liés entre eux et à différentes échelles. Notre objectif est donc de prendre en compte toutes ces notions : 1) considérer les dimensions temporelles et spatiales 2) généraliser le problème à une spatialisation plus complexe avec prise en compte des relations entre objets géographiques 3) extraire des motifs en considérant les différentes granularités possibles.

Dans la section 3.1, nous introduirons des définitions préliminaires qui seront la base de notre méthode. Nous présenterons ensuite la gestion des relations entre objets et la prise en compte des différentes granularités spatiales dans les sections 3.2 et 3.3. L'algorithme mis en place sera présenté dans la section 4. Dans la section 5, nous appliquerons notre méthode sur un jeu de données réelles et présenterons les résultats obtenus. Nous conclurons avec les perspectives envisagées dans le cadre de cette proposition dans la section 6.

3. Motif spatio-temporel relié

Notre approche est basée sur une extension des motifs séquentiels introduits par [Agrawal et Srikant1995] et prenant en compte la temporalité, et plus particulièrement des motifs spatio-séquentiels définis dans [Alatrista-Salas *et al.*2012] et prenant en compte la spatialité.

3.1. *Définitions préliminaires*

Les motifs séquentiels sont extraits à partir d'un ensemble de séquences de données. Pour chaque point ou objet géographique, une séquence d'événements est construite. Considérons une base de données $\mathcal{DB}$ comme celle présentée dans le tableau 1, regroupant l'ensemble des événements ayant eu lieu dans plusieurs villes. Chaque n-uplet correspond aux événements d'une ville à une date donnée et consiste en un triplet *(id-ville, id-date, item)* : l'identifiant de la ville, le mois ainsi que l'ensemble des événements (items) observés.

Ville	**Mois**	**Items**
Nîmes	01/2011	Humidité=Faible, Soleil
Montpellier	02/2011	Soleil
Nîmes	03/2011	Chaleur=Forte
Montpellier	03/2011	Humidité=Faible, Chaleur=Forte
Nîmes	04/2011	Chaleur=Faible, Vent
Orange	04/2011	Pluie
Orange	06/2011	Pluie, Vent

Tableau 1. *Base de données*

Pour chaque ville, nous générons sa séquence. L'ensemble S des séquences de l'exemple est donné par le tableau 2.

Définition 1 ***(Séquence)*** *Soit $\mathcal{I} = \{i_1, i_2, \ldots, i_m\}$ l'ensemble des $items$ (événements). Un itemset est un ensemble d'items non vide noté $(i_1, i_2, \ldots, i_k)$ où $i_j \in \mathcal{I}$ (il s'agit d'une représentation non ordonnée). Une séquence s est une liste ordonnée, non vide, d'itemsets notée $\langle is_1 is_2 \ldots is_p \rangle$ où $is_j \in \mathcal{IS}$, avec $\mathcal{IS}$ l'ensemble des itemsets.*

Chaque itemset étant composé d'items, nous pouvons considérer l'ensemble des items d'une séquence. Un item peut apparaître plusieurs fois dans une même séquence.

Ville	Séquence
Nîmes	⟨(Humidité=Faible, Soleil)(Chaleur=Forte)(Chaleur=Faible, Vent)⟩
Montpellier	⟨(Soleil)(Humidité=Faible, Chaleur=Forte)⟩
Orange	⟨(Pluie)(Pluie, Vent)⟩

Tableau 2. *Séquences d'évènements de villes*

L'extraction de connaissances à partir de séquences conduit à la recherche de sous-séquences fréquentes, également appelées motifs séquentiels. De nombreux algorithmes ont été proposés pour l'extraction de motifs classiques [Agrawal et Srikant1995, Ayres *et al.*2002, Zaki2001, Srikant et Agrawal1996, Masseglia *et al.*1998].

Définition 2 ***(Sous-séquence)*** *Une séquence $A = \langle is_1 is_2 \dots is_p \rangle$ est une sous-séquence d'une autre séquence $B = \langle is'_1 is'_2 \dots is'_m \rangle$ ($A \preceq B$) si $p \leq m$ et s'il existe des entiers $j_1 < j_2 < \dots < j_k < \dots < j_p$ tels que $is_1 \subseteq is'_{j_1}, is_2 \subseteq is'_{j_2}, \dots, is_p \subseteq is'_{j_p}$.*

Exemple 1 *Prenons les séquences du tableau 2. Chacune d'entre elles représente la séquence d'événements d'une ville. Nous pouvons voir que la séquence $s' = \langle$(Soleil)(Chaleur=Forte)$\rangle$ est supportée par les séquences des villes Nîmes et Montpellier. Donc nous avons $s' \preceq s_{N\hat{i}mes}$ et $s' \preceq s_{Montpellier}$.*

Un motif séquentiel est une sous-séquence fréquente caractérisée par un support. L'extraction de ces motifs est déterminée par un paramètre, le support minimum θ. Le support est le nombre d'occurrences du motif et seuls ceux avec un support supérieur au support minimum seront extraits. Soit $\mathcal{M}$ l'ensemble des motifs séquentiels extraits : $\forall m_i \in \mathcal{M}, Support(m_i) > \theta$.

Définition 3 ***(Support d'un motif séquentiel)*** *Une séquence $s \in S$ supporte un motif séquentiel m lorsque $m \preceq s$. Le support du motif m est le nombre de séquences de l'ensemble S qui supportent ce motif. Soit S_k l'ensemble des séquences supportant m, $S_k = \{s \in S \text{ tel que } m \preceq s\}$, $Support(m) = |S_k|$.*

Exemple 2 *En reprenant le tableau 2, nous voyons que la sous-séquence $s' = \langle$(Soleil)(Chaleur=Forte)$\rangle$ est supportée par les séquences des villes Nîmes et Montpellier. Nous trouvons alors $Support(s') = 2$.*

Même s'ils considèrent la temporalité, les motifs séquentiels ne gèrent pas la spatialité, ni les relations entre objets géographiques. Pour prendre en compte ces deux notions, nous allons présenter les notions de dimension et de hiérarchie sur une dimension.

3.2. *Relations entre objets spatiaux dans les motifs*

Les relations entre objets spatiaux sont les liens qui peuvent exister entre points ou objets géographiques. Par exemple dans un contexte épidémiologique, les objets spatiaux sont des villes ou régions. Plusieurs liens sont alors à considérer entre ces objets, comme les flux de déplacements ou une séparation symbolisée par la présence d'un obstacle, comme une montagne ou une forêt. Ces liens peuvent également se spécialiser. Il peut être intéressant de spécialiser le type d'obstacle par exemple, (cf. forêt de sapins ou forêt de chênes). Prendre en compte ces liens mais également pouvoir les spécialiser est donc nécessaire. Pour cela, nous allons utiliser un ensemble de dimensions de relations spatiales $\mathcal{D}_{\mathcal{R}}$ représentées par un ensemble de valeurs, avec pour chaque dimension, une hiérarchie associée.

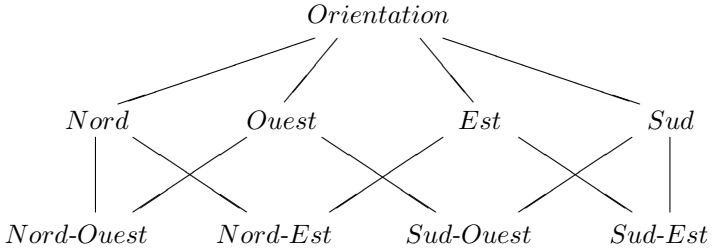
Définition 4 ***(Dimension de relations spatiales)*** *Une dimension $d \in \mathcal{D}_{\mathcal{R}}$ est définie par un domaine de valeurs x_j tel que $dom(d) = \{x_1, x_2, ..., x_n\}$.*

Exemple 3 *Dans une situation spatiale, soit $d_{Orientation}$ la dimension spatiale représentant l'orientation en fonction des points cardinaux.*
$dom(d_{Orientation}) = \{Nord, Ouest, Sud, Est\}$

Après avoir défini la notion de dimension, nous introduisons celle de hiérarchie sur une dimension. L'objectif est de pouvoir considérer une relation plus spécifique.

Définition 5 ***(Hiérarchie sur une dimension de relations spatiales)*** *Soit $d \in \mathcal{D}_{\mathcal{R}}$ une dimension de relations spatiales avec $dom(d) = \{x_1, x_2, ..., x_n\}$ et soit $h \in \mathcal{H}_{\mathcal{R}}$ la hiérarchie associée à cette dimension, alors h est un demi-treillis ou un arbre orienté et pour tout nœud $n \in h$, $label(n) \in dom(d)$.*

Exemple 4 *Soit $d_{Orientation}$ la dimension spatiale représentant l'orientation avec les points cardinaux, celle-ci est plus détaillée que dans l'exemple 3 avec les notions de Nord-Ouest, Nord-Est, Sud-Ouest et Sud-Est : $dom(d_{Orientation}) = \{Nord, Ouest, Sud, Est, Nord\text{-}Ouest, Nord\text{-}Est, Sud\text{-}Ouest, Sud\text{-}Est\}$. Nous pouvons alors construire la hiérarchie de relations suivante :*



Pour permettre la navigation dans cette hiérarchie, nous mettons en place différents opérateurs comme définis dans [Plantevit *et al.*2010]. Ce sont des opérations de généralisation et spécialisation directes et globales.

Définition 6 ***(Spécialisation directe et globale)*** *Soit $down_R(x_i)$ l'opération qui permet d'accéder aux spécialisations directes de la relation x_i et $downAll_R(x_i)$ l'opération qui permet d'accéder à toutes les spécialisations de la relation x_i. Les spécialisations directes de x_i sont tous les x_j tel qu'il y a un arc descendant de x_i à x_j dans la hiérarchie et les spécialisations globales de x_i sont tous les x_k tel qu'il y a un chemin descendant de x_i à x_k.*

Exemple 5 *Prenons la dimension $d_{Orientation}$:*
- $down_R(Ouest) = \{Nord\text{-}Ouest, Sud\text{-}Ouest\}$
- $downAll_R(Orientation) = dom(d_{Orientation})$

Définition 7 ***(Généralisation directe et globale)*** *Soit $up_R(x_i)$ l'opération qui permet d'accéder aux généralisations directes de la relation x_i et $upAll_R(x_i)$ l'opération qui permet d'accéder à toutes les généralisations de la relation x_i Les généralisations directes de x_i sont tous les x_j tel qu'il y a un arc ascendant de x_i à x_j dans la hiérarchie et les généralisations globales de x_i sont tous les x_k tel qu'il y a un chemin ascendant de x_i à x_k.*

Exemple 6 *Reprenons l'exemple précédent :*
- $up_R(Nord\text{-}Est) = \{Nord, Est\}$
- $upAll_R(Nord\text{-}Ouest) = \{Nord, Ouest, Orientation\}$.

En construisant cette hiérarchie, nous offrons la possibilité d'extraire l'information à plusieurs niveaux. Nous pouvons par exemple mettre en évidence l'importance de la présence d'un événement au Nord, mais il est également possible de descendre dans la hiérarchie et de trouver des relations plus spécifiques. Après avoir défini les hiérarchies sur des dimensions et les opérations, notre but est de nous en servir dans l'extraction de motifs qui prennent en compte les relations entre objets spatiaux. Comme dans [Alatrista-Salas *et al.*2012] nous utilisons l'opérateur de relation **.** et l'opérateur de groupement **[]** dans les motifs séquentiels. La différence vient de l'association d'une hiérarchie avec une ou plusieurs dimensions. Ces motifs sont alors constitués par une séquence d'itemsets reliés.

Définition 8 ***(Itemset relié)***
*Soient deux itemsets is_i , is_j, s'il existe un lien δ, de la hiérarchie d'une dimension de relations entre objets géographiques, entre is_i et is_j, alors ils formeront un itemset relié, noté $is_R = is_i\ .\delta[is_j]$. L'opérateur de liaison **.** est suivi du nom de la relation où **[]** est un opérateur de groupement : ainsi si deux itemsets is_k, is_l sont liés à is_i par la même relation, on notera $is_i\ .\delta[is_k is_l]$ ou de manière équivalente $is_i\ .\delta[is_l is_k]$. Si is_i est lié à des itemsets par plusieurs relations, on notera indifféremment $is_i.\delta[is_k].\gamma[is_l]$ ou $is_i.\gamma[is_l].\delta[is_k]$.*

Exemple 7 *Avec une relation de voisinage classique, nous pouvons avoir le motif* $\langle$(*Humidité .[Pluie Vent]*)(*Humidité Pluie*)$\rangle$. *Imaginons que nous ayons mis en évidence que la pluie et le vent soient apparus à chaque fois dans un endroit voisin situé au nord, nous trouverions alors le motif :* $\langle$(*Humidité* $._{Nord}$*[Pluie Vent]*)(*Humidité Pluie*)$\rangle$

En utilisant la hiérarchie, nous extrayons tous les motifs, des plus généraux aux plus spécifiques. Nous définissons maintenant l'inclusion d'un motif spatio-temporel relié dans un autre motif spatio-temporel relié. Cette définition est la même que celle de la séquence classique, seule l'inclusion d'un itemset dans un autre est différente.

Définition 9 ***(Inclusion d'un itemset relié)***
Un itemset relié $is_R = is_i.\delta[is_j]$ *est inclus dans un autre itemset relié* $is'_R = is'_i.\delta'[is'_j]$, *si et seulement si* $is_i \subseteq is'_i$, $is_j \subseteq is'_j$ *et* $\delta = \delta'$ *ou* $\delta \in upAll(\delta')$ *(i.e.* δ' *est égal à* δ *ou alors* δ' *est une spécialisation de* δ*).*

Exemple 8 *Soit* $d_{Orientation}$ *la dimension relationnelle de deux itemsets* is_1 *et* is_2 *tels que* $is_1 = ._{Sud}$[*Humidité*,*Vent*] *et* $is_2 = ._{Est}$[*Humidité*,*Pluie*,*Vent*]. *On voit que tous les items de* is_1 *sont inclus dans* is_2 *et que la relation de* is_1 *est plus générale dans la hiérarchie que* is_2. *Donc* $is_1 \preceq is_2$. *Nous pouvons trouver un motif avec l'itemset* $is' = ._{Sud}$[*Humidité*,*Vent*]$._{Nord}$[*Pluie*]. *Il s'interprète de la manière suivante : les événements Humidité et Vent ont eu lieu dans une ville au sud et, au même moment, l'événement Pluie à eu lieu dans une ville au nord.*

Les motifs séquentiels obtenus sont composés d'itemsets reliés pour former un nouveau type de motif, les motifs séquentiels reliés.

Définition 10 ***(Motif séquentiel relié)***
Soit $\mathcal{IS}$ *l'ensemble des itemsets et* $\mathcal{IS}_\mathcal{R}$ *l'ensemble des itemsets reliés, un motif séquentiel relié* m_R *est une liste ordonnée, non vide, d'itemsets et d'itemsets reliés notée* $\langle is_1, is_2, \ldots, is_p\rangle$ *où* $is_j \in \mathcal{IS} \cup \mathcal{IS}_\mathcal{R}$ *avec une valeur de support* $Support(m_R)$.

Dans cette section, nous avons présenté un nouveau type de motifs séquentiels qui considère les relations pouvant exister entre les objets géographiques où se déroulent des événements. Ces relations peuvent être hiérarchisées. Mais dans un contexte où nous avons un découpage spatial, il est également intéressant d'ajouter l'information spatiale au motif, comme nous allons le présenter dans la section suivante.

3.3. *Granularités géographiques dans les motifs*

Les différentes granularités géographiques sont le plus souvent représentées par un découpage de l'espace. Ce découpage peut prendre différentes formes en fonction du

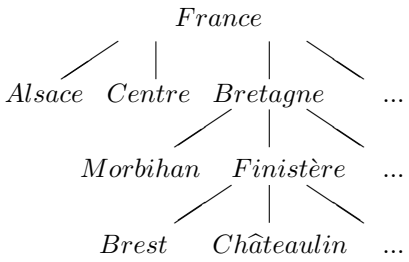
contexte et du problème à résoudre. Par exemple, si nous envisageons un découpage de la Terre d'un point de vue géopolitique, alors l'espace est divisé en fonction des frontières, des continents ou pays. D'un point de vue climatique, ce découpage est différent avec des zones à climat chaud, zones à climat tempéré, etc. De plus ces zones sont elles mêmes divisées en zones de plus petite taille. Un continent est divisé en pays et une zone à climat tempéré est divisée en zones à climat continental, océanique, etc. Il est donc important de prendre d'une part en compte ce découpage en zones mais également de considérer un découpage plus spécifique de ces zones.

Définition 11 ***(Dimension de zones spatiales)*** *Une dimension de zones spatiales $d \in \mathcal{D}_{\mathcal{S}}$ est défini par un domaine de valeurs x_j tel que $dom(d) = \{x_1, x_2, ..., x_n\}$.*

Exemple 9 *Soit $d_{Département} \in \mathcal{D}_{\mathcal{S}}$ la dimension de zones spatiales représentant le découpage administratif de la France en départements.*
$dom(d_{Département}) = \{Ain, Aisne, Allier, ..., Val\text{-}de\text{-}Marne, Val\text{-}d'Oise\}$

Une prise en compte de la granularité implique la construction d'une hiérarchie basée sur une relation d'inclusion à partir du découpage géographique. Pour illustrer notre approche, nous allons reprendre comme exemple le découpage administratif de la France mais en considérant maintenant les différentes granularités. Celle-ci est divisée en régions, elles-mêmes divisées en départements, ensuite en arrondissements. Les cantons et communes ne seront pas considérés dans cet exemple. La hiérarchie suivante représentent ce découpage.

Exemple 10 *Hiérarchie du découpage administratif de la France*



La représentation sous forme hiérarchique des granularités spatiales est proche des hiérarchies utilisées dans le cadre des relations entre objets spatiaux. Néanmoins, cette hiérarchie n'est pas basée sur les principes de généralisation/spécialisation mais sur la notion d'inclusion. Par exemple, un département n'est pas une spécialisation d'une région mais une sous-division. Nous devons alors redéfinir les opérations de parcours dans la hiérarchie $h \in \mathcal{H}_{\mathcal{S}}$, avec $\mathcal{H}_{\mathcal{S}}$ l'ensemble des hiérarchies sur les dimensions de zones spatiales.

Définition 12 ***(Contenu direct et global)*** *Soit $down_S(x_i)$ l'opération qui permet d'accéder aux contenus directs de la valeur de granularité x_i et $downAll_S(x_i)$ l'opération qui permet d'accéder à tous les contenus de la valeur de granularité x_i. Les contenus directs de x_i sont tous les x_j tel qu'il y a un arc descendant de x_i à x_j dans la hiérarchie et les spécialisations globales de x_i sont tous les x_k tel qu'il y a un chemin descendant de x_i à x_k.*

Définition 13 ***(Contenant direct et global)*** *Soit $up_S(x_i)$ l'opération qui permet d'accéder aux contenants directs de la valeur de granularité x_i et $upAll_S(x_i)$ l'opération qui permet d'accéder à toutes les contenants de la valeur de granularité x_i. Les contenants directs de x_i sont tous les x_j tel qu'il y a un arc ascendant de x_i à x_j dans la hiérarchie et les contenants globaux de x_i sont tous les x_k tel qu'il y a un chemin ascendant de x_i à x_k.*

À partir de ce nouveau type de hiérarchie, nous pouvons rajouter une information spatiale aux motifs obtenus. Certains motifs seront fréquents dans les villes de la France entière mais il se peut que certains motifs soient fréquents sur une région, un département ou bien un arrondissement. Pour extraire ce type de motifs, nous parcourons alors toutes les hiérarchies de granularités et vérifions si un motif est également fréquent à un niveau plus spécifique des hiérarchies. Si tel est le cas, le motif devient alors spatio-temporel car sa fréquence est dépendante d'une zone spatiale.

Définition 14 ***(Motif spatio-temporel)***
Soit [] l'opérateur de spatialité et m un motif séquentiel classique ou sequentiel-relié, $x_k \in d$ la valeur d'une dimension de zones spatiales d, un support minimum θ et $\mathcal{S}'$ l'ensemble des séquences s_i tel que $|m \preceq s_i|$ à la valeur de granularité x_k. Si $|\mathcal{S}'| > \theta$ alors un motif spatio-temporel m' tel que $m' = [x_k]m$ est créé.

Exemple 11 *Soit le motif séquentiel relié $m' = \langle$(Humidité $._{Nord}$[Pluie Vent]) (Humidité Pluie)$\rangle$, avec $\theta = 10\%$ et $Support(m') = 50\%$.*
Le motif m' a une fréquence de 50% dans la France entière. Mais sa fréquence est de 15% si nous ne considérons que les villes de la région Alsace. Un motif spatio-temporel m'' est alors créé tel que $m'' = [Alsace]\langle$(Humidité $._{Nord}$[Pluie Vent])(Humidité Pluie)$\rangle$ et $Support(m'') = 15\%$. Ici la relation .Nord désigne des évènements dans des villes au nord des villes alsaciennes, et non nécessairement en Alsace.

Les définitions présentées précédemment permettent la prise en compte des relations spatiales et des granularités géographiques. Leur utilisation dans un algorithme adapté permet l'extraction de motifs spatio-temporels reliés, et à différentes échelles (i.e. à différentes granularités spatiales). L'algorithme qui a été mis en place est présenté dans la section suivante.

4. Algorithme STR_PrefixGrowth

Pour extraire les motifs, nous nous sommes appuyés sur l'algorithme d'extraction PrefixSpan[Pei *et al.*2004], également utilisé dans [Alatrista-Salas *et al.*2012]. Il est actuellement l'un des algorithmes les plus performants pour extraire des motifs séquentiels, que ce soit en terme de temps de calcul ou bien en consommation mémoire. Les motifs séquentiels sont extraits à partir de la notion de préfixes fréquents. Par exemple, $\langle(a)\rangle$, $\langle(a)(a)\rangle$, $\langle(a)(ab)\rangle$ et $\langle(a)(abc)\rangle$ sont préfixes de la séquence $\langle(a)(abc)(ac)(d)(cf)\rangle$. Si un préfixe est présent dans un nombre de séquences supérieur à un support minimum θ, alors ce préfixe est considéré comme fréquent. Lorsqu'un préfixe fréquent ou plusieurs sont trouvés, la base de données est alors divisée de manière récursive. En effet, lorsque nous cherchons les motifs fréquents, il n'est pas nécessaire de conserver toute la base de données. Les données (i.e. séquences) qui ne supportent pas le motif courant ne sont pas conservées dans la base projetée. La raison est que ces séquences ne supporteront pas non plus les motifs de longueur supérieure à cause de la propriété antimonotonique du support. L'efficacité de cet algorithme est due à (1) la non génération de motifs candidats grâce à la recherche des préfixes fréquents (2) la projection de la base de données en plus petites bases qui permet d'accélérer l'exploration en enlevant les séquences qui ne sont plus nécessaires.

Algorithme 1: $STR\_PrefixGrowth(\alpha, \theta, \mathcal{BD}|_\alpha, \mathcal{D}_\mathcal{R}, \mathcal{D}_\mathcal{S})$

input : α un motif spatio-temporel, θ le support minimum, $\mathcal{BD}|_\alpha$ la base de données projetée en fonction du motif α , $\mathcal{D}_\mathcal{R}$ l'ensemble des hiérarchies sur les dimensions concernant les relations , $\mathcal{D}_\mathcal{S}$ l'ensemble des hiérarchies sur les dimensions spatiales

output : $\mathcal{MS}$ l'ensemble des motifs extraits dans cet appel de fonction (i.e récursion courante)

$I_\theta \leftarrow$ getListOccurences($\theta, \mathcal{BD}|_\alpha, \mathcal{D}_\mathcal{R}$);
$\mathcal{MS} \leftarrow \varnothing$;
foreach *i in* I_θ **do**
  β=append(α,i);
  $\mathcal{MS} \leftarrow \mathcal{MS} \cup \beta$;
  $\mathcal{MS} \leftarrow \mathcal{MS} \cup$ STR_PrefixGrowth($\beta, \theta, \mathcal{BD}|_\beta, \mathcal{D}_\mathcal{R}, \mathcal{D}_\mathcal{S}$);
  $\mathcal{MS} \leftarrow \mathcal{MS} \cup$ explorerSpatialHierarchy($\beta, \theta, \mathcal{BD}|_\beta, \mathcal{D}_\mathcal{S}$);
end

Notre approche générale est décrite par l'algorithme récursif 1 que nous appellerons $STR\_PrefixGrowth$ pour **S**patio-**T**emporal and **R**elational **PrefixGrowth**. Cette méthode récupère tout d'abord la liste des occurrences fréquentes dans la base projetée sur α en fonction du support minimum θ. Une occurrence fréquente (e.g item fréquent) signifie qu'un motif de longueur supérieure a été trouvé. C'est dans la fonction $getListOccurences()$, que nous explorons les hiérarchies de relations. Deux opérations sont mises en jeu, $searchIExtend()$ et $searchSExtend()$ qui représentent les deux manières d'étendre un motif, la I-Extension et la S-Extension. La I-Extension ajoute un item dans le dernier itemset de la séquence et la S-Extension

Algorithme 2: $getListOccurences(\theta, \mathcal{BD}|_\alpha, \mathcal{D}_\mathcal{R})$

input : θ le support minimum, $\mathcal{BD}|_\alpha$ la base de données projetée en fonction d'un motif α , $\mathcal{D}_\mathcal{R}$ l'ensemble des hiérarchies sur les dimensions concernant les relations
output : I_θ la liste des occurrences fréquentes dans $\mathcal{BD}|_\alpha$
$I_\theta \leftarrow I_\theta \cup$ searchIExtend(θ,$\mathcal{BD}|_\alpha$);
$I_\theta \leftarrow I_\theta \cup$ searchSExtend(θ,$\mathcal{BD}|_\alpha$);
foreach dim_i *in* $\mathcal{D}_\mathcal{R}$ **do** `/* Pour chaque dimension dans` $\mathcal{D}_\mathcal{R}$ `*/`
  $I_\theta \leftarrow I_\theta \cup$ searchIntend(θ,$\mathcal{BD}|_\alpha$,dim_i);
  $I_\theta \leftarrow I_\theta \cup$ searchExtend(θ,$\mathcal{BD}|_\alpha$,dim_i);
end

Algorithme 3: $exploreSpatialHierarchy(\alpha, \theta, \mathcal{BD}|_\alpha, \mathcal{D}_\mathcal{S})$

input : α un motif spatio-temporel et multi-niveaux , θ le support minimum, $\mathcal{BD}|_\alpha$ la base de données projetée en fonction d'un motif α , $\mathcal{D}_\mathcal{S}$ l'ensemble des hiérarchies sur les dimensions spatiales
output : $\mathcal{MS}$ l'ensemble des motifs extraits
$\mathcal{MS} \leftarrow \varnothing$;
foreach dim_i *in* $\mathcal{D}_\mathcal{S}$ **do**
  foreach s *in* dim_i **do**
    if *isFrequent*($\alpha, \theta, \mathcal{BD}|_\alpha$,*s)* **then** `/* teste si le motif est fréquent dans la granularité spatiale courante */`
      $\mathcal{MS} \leftarrow \mathcal{MS} \cup$ spatialPattern(α, s);
  end
end

ajoute un item dans un nouvel itemset à la fin de la séquence, à une date différente. Par exemple prenons le motif $m = \langle(a)(b)\rangle$ et une occurrence fréquente représentant l'item c. Si c est une I-Extension et m' le motif étendu, alors $m' = \langle(a)(bc)\rangle$. Si c est une S-Extension et m'' le motif étendu, alors $m'' = \langle(a)(b)(c)\rangle$. Pour chaque hiérarchie de relation, les opérations $searchIExtend()$ et $searchSExtend()$ sont appliquées pour trouver les occurrences des relations à chaque étage de la hiérarchie. Les relations fréquentes sont alors considérées sous forme d'occurrences. Les relations entre les séquences étant gérées comme des items particuliers, elles seront retournées en même temps que les occurrences des items classiques. Cette fonction est présentée par l'algorithme 2. Les occurrences, ou items fréquents, vont servir alors à étendre le motif α par la fonction $append()$ qui prend en compte le fait que l'item soit une intention ou une extension. Ensuite, pour chaque motif étendu β, nous projetons la base de données en fonction de ce motif et nous appelons alors $STR\_PrefixGrowth$ pour continuer la fouille récursive des motifs. Pour finir, chaque motif est donné en pa-

ramètre de la fonction $exploreSpatialHierarchy()$ qui va explorer les dimensions spatiales à tous les niveaux de granularité (algorithme 3) pour trouver de nouveaux motifs (section 3.3) sur la spatialité. Pour chaque dimension spatiale, nous vérifions si le motif est fréquent à chaque granularité de la hiérarchie. Si c'est le cas, nous ajoutons le motif spatial à l'ensemble des motifs.

La complexité dans le pire des cas de PrefixSpan pour l'extraction des motifs séquentiels est $\Theta((2 \cdot I)^L)$ avec I le nombre d'items et L la longueur de la plus grande séquence de la base de données $\mathcal{BD}$. Soit H_R le nombre de hiérarchies de relations spatiales, soit R le nombre de relations maximale par hiérarchie de relations spatiales, soit H_S le nombre de hiérarchies de granularité et soit S le nombre de zone spatiale maximale par hiérarchie de granularité, alors la complexité dans le pire des cas de l'algorithme STR_PrefixGrowth est $\Theta(H_S \cdot S \cdot (2 \cdot N \cdot H_R \cdot R)^L)$. Cet algorithme est pseudo-polynomial, c'est à dire linéaire en fonction du nombre de motifs extraits. Le pire des cas correspond donc au nombre maximal de motifs qui peuvent être extraits à partir d'un jeu de données.

Pour tester et valider notre méthode, nous avons appliqué cet algorithme sur un jeu de données réel et nous l'avons comparé à des méthodes existantes. Ces résultats sont présentés dans la section suivante.

5. Application aux données hydrologiques

Les données que nous étudions ont été produites par l'agence de l'Eau Rhône-Méditerranée et Corse et sont mises à notre disposition dans le cadre du projet de l'ANR Fresqueau. Elles renseignent sur les caractéristiques physico-chimiques et biologiques des cours d'eaux du bassin versant de la Saône, bassin qui s'étend sur tout ou partie de 11 départements de l'est de la France. Les données ont été collectées à différentes dates, sur plusieurs stations (711) le long de ces cours d'eau. Elles sont de différentes formes : mesures (ex : température de l'eau, pH, taux de nitrates, oxygène dissous, etc.), indices (ex : indice biologique IBGN), etc. Pour chaque station, les données collectées aux différentes dates constituent des itemsets qui sont ordonnés pour générer une séquence. De plus, pour appliquer notre méthode, nous sélectionnons plusieurs caractéristiques qui vont servir à expliciter la spatialité des données et à prendre en compte les différentes échelles géographiques.

5.1. *Hiérarchies de relations spatiales*

Ces données sont représentées par une dimension avec sa hiérarchie associée, pour prendre en compte les granularités et les relations entre stations. Trois types d'informations ont été étudiées et sont présentées si-dessous.

Orientation des cours d'eau : elle permet de savoir si une station se situe en aval ou en amont d'une autre station. C'est donc une hiérarchie simple avec un seul niveau de profondeur.

Les aires hydrographiques : la France a été découpée en bassins versants, soit quatre partitions hiérarchisées. Chaque niveau est un découpage plus spécifique du niveau précédent. Au niveau le plus général nous avons les bassins hydrographiques, eux mêmes découpés en région hydrographique. Ensuite nous avons les secteurs divisés en sous-secteurs. Nous obtenons alors une hiérarchie de profondeur 4.

Chaque station est soit en aval, soit en amont d'une station voisine et est associée à une aire hydrographique. L'orientation des cours d'eau est utilisée comme dimension de relation entre les stations (section 3.2). L'aire hydrographique sert à prendre en compte la granularité géographique (section 3.3).

5.2. *Expérimentations*

L'extraction de motifs nécessite au préalable une discrétisation des données. Nous avons arbitrairement choisi une discrétisation avec 5 intervalles pour chaque type de données. Nous donnons ici une description des données qui apparaissent dans les motifs présentés dans le tableau 4 :

ibgn : L'indice biologique global normalisé (IBGN) est un outil utilisé pour l'évaluation de la qualité biologique d'un cours d'eau. Cet indice prend une valeur entre 0 et 20 en fonction de la présence ou non de certains bio-indicateurs.

ibgn_note : C'est une note variant de 0 à 5 basée sur la valeur de l'indice IBGN.

var_taxo : Cette donnée décrit la variété taxonomique. C'est une métrique qui correspond au nombre de taxons (macroinvertébrés d'eau douce) récoltés au cours d'un prélèvement.

θ	**MS**	$\mathbf{M}_{ST}$	$\mathbf{M}_{STR}$
0.5	1	4	4
0.4	4	12	12
0.3	22	60	64
0.2	75	186	233
0.1	180	445	1882

Tableau 3. *Nombre de motifs extraits en fonction du support minimum*

Nous comparons notre approche avec l'extraction de motifs séquentiels classiques (**MS**) et spatio-temporels obtenus avec la méthode de [Alatrista-Salas *et al.*2012] ($\mathbf{M}_{ST}$). Ces deux méthodes sont les plus proches de la nôtre, que nous appellerons $\mathbf{M}_{STR}$ pour motifs **S**patio-**T**emporels et **R**eliés. Dans le tableau 3, nous avons fait varier la valeur du support minimum pour observer l'évolution du nombre de motifs extraits en fonction des différentes méthodes. Le tableau 4 présente un échantillon de motifs extraits avec chacune des méthodes. L'ajout de différentes relations spatiales ainsi que la navigation dans les hiérarchies permettent l'extraction de motifs plus spécifiques et plus expressifs, que nous n'aurions jamais pu extraire avec une méthode

Méthode	Séquence	Support
MS	$\langle(var\_taxo\_31\text{-}40)\rangle$	0.404
$\mathbf{M}_{ST}$	$\langle(.[ibgn\_11\text{-}15])(var\_taxo\_31\text{-}40)\rangle$	0.089
$\mathbf{M}_{STR}$	$\langle(.Orient[ibgn\_11\text{-}15])(var\_taxo\_31\text{-}40)\rangle$	0.089
	$\langle(.Aval[ibgn\_11\text{-}15])(var\_taxo\_31\text{-}40)\rangle$	0.051
	[U1]$\langle(.Orient[ibgn\_11\text{-}15])(var\_taxo\_31\text{-}40)\rangle$	0.054
	[U2]$\langle(.Orient[ibgn\_11\text{-}15])(var\_taxo\_31\text{-}40)\rangle$	0.073

Tableau 4. *Motifs obtenus avec les différentes méthodes*

classique. Par exemple, le motif [U2]$\langle(.Orient[ibgn\_11\text{-}15])(var\_taxo\_31\text{-}40)\rangle$ (tableau 4) se lit de la manière suivante : dans le bassin hydrographique U2 nous trouvons fréquemment une valeur IBGN comprise entre 11 et 15 dans une station voisine (i.e. en amont **ou** en aval) et plus tard dans le temps une variété taxonomique comprise entre 31 et 40. Ce motif ne peut pas être obtenu avec les motifs séquentiels classiques, cf. $\langle(var\_taxo\_31\text{-}40)\rangle$, ni avec la méthode de [Alatrista-Salas *et al.*2012], cf. $\langle(.[ibgn\_11\text{-}15])(var\_taxo\_31\text{-}40)\rangle$. Il est souvent difficile pour les experts de déterminer la meilleure échelle, c'est-à-dire celle qui permet d'obtenir les meilleures observations et il est possible que les paramètres ne soient pas tous observables à la même granularité. Notre approche autorise la présence de différents niveaux d'une hiérarchie dans les résultats. Les motifs extraits répondent ainsi à plusieurs problématiques : 1) la prise en compte des dimensions spatiales et temporelles 2) la gestion des relations entre les objets 3) l'exploration et la mise en évidence de la granularité la plus adaptée.

6. Conclusion et perspectives

Nous avons proposé une méthode capable de répondre à de nouveaux besoins. Les motifs extraits permettent de gérer de manière efficace les dimensions spatiales et temporelles. Notre approche se démarque des solutions proposées dans la littérature, par une gestion plus fine de la spatialité avec les notions de relations spatiales avec différentes granularités. Le résultat est l'obtention de motifs plus riches sémantiquement. Ce type d'extraction entraîne l'exploration d'un immense espace de recherche et l'obtention d'un nombre important de motifs. Dans la poursuite de ce travail, nous souhaitons dans un premier temps nous orienter vers l'application de mesures d'élagage durant l'extraction des motifs, pour accélérer l'extraction et limiter le nombre de motifs. Et dans un second temps, proposer des mesures d'intérêts aux experts pour filtrer cette connaissance selon des critères spécifiques.

Ces travaux ont été financés dans le cadre du projet Fresqueau (ANR11_MONU14).

7. Bibliographie

Agrawal R., Srikant R., « Mining Sequential Patterns », *Proceedings of the Eleventh International Conference on Data Engineering*, ICDE '95, Washington, DC, USA, 1995, p. 3–14.

Alatrista-Salas H., Bringay S., Flouvat F., Selmaoui-Folcher N., M.Teisseire, « Vers une approche efficace d'extraction de motifs spatio-séquentiels », *EGC*, 2012.

Ayres J., Flannick J., Gehrke J., Yiu T., « Sequential PAttern mining using a bitmap representation », *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, New York, NY, USA, 2002, ACM, p. 429–435.

Gubler D. J., « Epidemic dengue/dengue hemorrhagic fever as a public health, social and economic problem in the 21st century. », *Trends in Microbiology*, vol. 10, n° 2, 2002, p. 100–103.

Huang Y., Zhang L., Zhang P., « A Framework for Mining Sequential Patterns from Spatio-Temporal Event Data Sets », *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, n° 4, 2008, p. 433–448.

Lin X., Mei Q., Han J., Jiang Y., , Danilevsky M., « Inferring the Diffusion and Evolution of Topics in Social Communities », *Evolution*, vol. 3, n° 3, 2011, p. 1231–1240.

Masseglia F., Cathala F., Poncelet P., « The PSP Approach for Mining Sequential Patterns », 1998, p. 176–184.

Pei J., Han J., Member S., Mortazavi-asl B., Wang J., Pinto H., Chen Q., Dayal U., Society I. C., Society I. C., chun Hsu M., « Mining Sequential Patterns by Pattern-Growth : The PrefixSpan Approach », *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, 2004, page 2004.

Plantevit M., Laurent A., Laurent D., Teisseire M., Choong Y. W., « Mining multidimensional and multilevel sequential patterns », *ACM Trans. Knowl. Discov. Data*, vol. 4, n° 1, 2010, p. 4 :1–4 :37, ACM.

Srikant R., Agrawal R., « Mining Sequential Patterns : Generalizations and Performance Improvements », , 1996, p. 3–17.

Tsoukatos I., Gunopulos D., « Efficient Mining of Spatiotemporal Patterns », *Advances in Spatial and Temporal Databases*, vol. 2121, 2001, p. 425-442.

Wang J., Hsu W., Lee M., « Mining Generalized Spatio-Temporal Patterns », *Database Systems for Advanced Applications*, vol. 3453 de *Lecture Notes in Computer Science*, 2005, p. 989-989.

Zaki M. J., « SPADE : An Efficient Algorithm for Mining Frequent Sequences », *Mach. Learn.*, vol. 42, n° 1-2, 2001, p. 31–60.