Motifs spatio-temporels. Enjeux et applications à l'environnement

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

Motifs spatio-temporels

Enjeux et applications à l'environnement

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

1. *PPME, Université de la Nouvelle Calédonie, BP R4 98851 Nouméa, Nouvelle Calédonie*
   {nazha.selmaoui,frederic.flouvat,hugo.alatrista}@univ-nc.nc
2. *Irstea - TETIS, 500, rue J. F. Breton 34093 Montpellier, France*
   hugo.alatrista-salas@teledetection.fr
3. *LIRMM, 161 rue Ada, 34095 Montpellier Cedex 5, France*
   bringay@lirmm.fr

RÉSUMÉ. *Les avancées technologiques en termes d'acquisition de données permettent de mieux surveiller les phénomènes évolutifs dans divers domaines dont l'environnement. Les données collectées sont de plus en plus complexes (spatiales, temporelles, hétérogènes et multi-échelles). L'exploitation de ces données par leurs propriétaires (experts du domaine) nécessite de concevoir de nouvelles méthodes d'analyse de données et de découverte de connaissances. Dans ce contexte, les approches de découverte de motifs spatio-temporels se révèlent particulièrement pertinentes. Ce papier propose de faire une revue détaillée de ces travaux. Nous nous focalisons sur deux exemples de motifs : les colocalisations et les motifs spatio-séquentiels. Ces motifs ont été utilisés pour étudier deux applications réelles dans le domaine de l'environnement.*

ABSTRACT. *Technological advances in terms of data acquisition enable to better monitor dynamic phenomena in various domains including environment. The collected data are more and more complex (spatial, temporal, heterogeneous and multi-scale). The exploitation of this data requires new methods of data analysis and knowledge discovery. In this context, approaches for discovering spatio-temporal patterns are particularly relevant. This paper proposes to make a detailed review of these works. We focus on two examples of patterns : colocation and spatio-sequential patterns. These patterns have been used to study real applications in the field of environment.*

MOTS-CLÉS : *fouille de données spatio-temporelles, colocalisations, motifs séquentiels.*

KEYWORDS: *spatio-temporal data mining, colocations, sequential patterns.*

DOI:10.3166/RIA.27.619-648

1. Introduction

La fouille de données (ou extraction de connaissances) est définie comme "le processus non trivial d'extraction d'informations implicites, nouvelles, et potentiellement utiles à partir de (grands volumes de) données" (Piatetsky-Shapiro, 1991). C'est un domaine qui connaît une croissance assez spectaculaire, sous l'impulsion des organisations propriétaires de grands volumes de données et soucieuses d'en extraire de la valeur ajoutée. La fouille de données consiste à induire des lois générales à partir des données stockées. Ces généralisations sont le plus souvent des énoncés de haut niveau, comme par exemple des règles descriptives ou des arbres de décision. La découverte de motifs (*Pattern Discovery*) dans les données est l'un des problèmes phare en fouille de données. Il a été beaucoup étudié en bioinformatique pour l'analyse des données génomiques à large échelle et dans de nombreux domaines d'applications où des régularités peuvent être porteuses de valeur ajoutée, comme par exemple, la découverte de règles d'associations dans des données transactionnelles (Agrawal *et al.*, 1993), de motifs séquentiels dans des bases de séquences de comportements (Agrawal, Srikant, 1995 ; Mannila *et al.*, 1997 ; Masseglia *et al.*, 1998) ou découverte de motifs plus complexes tels que des sous-graphes (Inokuchi *et al.*, 2000) ou des sous-arbres (Termier *et al.*, 2002 ; M. J. Zaki, 2002).

Ces dernières années, les avancées technologiques en termes d'acquisition des données (images satellitaires [1], capteurs, etc.) ont donné lieu à un grand nombre d'applications dans la surveillance et le suivi environnemental tels que la détection de changements abrupts (catastrophes naturelles...), le suivi de phénomènes évolutifs (érosion côtière, désertification, feux de brousse...) ou la mise au point de modèles (hydrologie, activité agricole...). Les données collectées sont généralement hétérogènes, multi-échelles, spatiales et temporelles (série temporelle d'images satellites, aériennes ou terrestres ; modèles numériques de terrain, mesures physiques au sol, observations qualitatives...) et sont destinées à comprendre et prédire des phénomènes résultant de processus complexes et d'origine pluridisciplinaire (données climatiques, géologiques...). L'exploitation de cette masse de données par les experts nécessite non seulement de les structurer au mieux mais aussi et surtout de concevoir des méthodes d'analyse de données et de découverte de connaissances. Dans ce contexte, les approches de découverte de motifs se révèlent très pertinentes.

Face à l'explosion de l'information spatiale et des systèmes d'information géographique (SIG), de nombreux travaux ont été réalisés dans le contexte de l'extraction de motifs spatio-temporels. Les premiers travaux dans ce domaine ont traité les dimensions spatiale et temporelle séparément. L'extraction de séries temporelles vise à repérer des caractéristiques fréquentes dans le temps (Agrawal, Srikant, 1995 ; Srikant, Agrawal, 1996 ; Mannila *et al.*, 1997 ; M. Zaki, 2001 ; Pei *et al.*, 2004) sans prendre en compte les relations spatiales. La recherche de colocalisations (Huang *et*

1. Les capacités des satellites d'observation de la Terre se sont considérablement accrues dans les trois composantes que sont la fréquence d'acquisition (essentiellement liée au nombre de capteurs disponibles), la résolution spatiale (50 cm) et la discrimination spectrale (Moyen Infra-Rouge et bande bleue en standard).

al., 2004 ; Zhang *et al.*, 2004 ; Huang *et al.*, 2006 ; Yoo, Shekhar, 2006 ; Verhein, Al-Naymat, 2007 ; Yoo, Bow, 2009) extrait un ensemble de caractéristiques apparaissant fréquemment dans des objets voisins sans prendre en compte la temporalité. Plus récemment, ces travaux ont été étendus pour intégrer simultanément les dimensions spatiale et temporelle. Les travaux sur l'identification de séquences d'évènements localisés (Tsoukatos, Gunopulos, 2001 ; J. Wang *et al.*, 2004 ; 2005 ; Celik *et al.*, 2006 ; 2008 ; Huang *et al.*, 2008) ou la détection de trajectoires (Mamoulis *et al.*, 2004 ; Cao *et al.*, 2005 ; Yang *et al.*, 2005 ; Giannotti *et al.*, 2007 ; Qian *et al.*, 2009) en sont deux exemples. Une synthèse récente vient d'être éditée par le consortium GeoPKDD (Giannotti, Pedreschi, 2008). Cependant, dans ces travaux, les domaines de motifs utilisés et donc les motifs extraits ne sont pas à la hauteur de la complexité spatiale des objets à étudier dans des images satellitaires. De même, les contraintes primitives généralement étudiées (typiquement des fréquences minimales) restent insuffisantes pour caractériser les critères d'intérêt des experts comme les géologues.

Dans cet article, nous nous intéressons aux problèmes de l'extraction de motifs décrivant, en plus des relations thématiques (attributs), des relations spatiales et/ou temporelles. Dans une première partie, nous présentons un état de l'art détaillé des méthodes permettant d'exploiter les données spatiales et spatio-temporelles. Nous décrivons ensuite deux méthodes d'extraction de motifs en lien avec des applications réelles. Les premiers types de motifs étudiés sont des motifs uniquement spatiaux : les colocalisations. Ils sont extraits dans des systèmes d'information géographique et ont pour but d'étudier et de caractériser l'érosion des sols. Les deuxièmes types de motifs étudiés sont des motifs intégrant les dimensions spatiale et temporelle : les motifs spatio-séquentiels. Ces motifs sont appliqués à deux jeux de données : des données épidémiologiques et des données sur l'érosion des sols. Nous concluons ensuite et présentons les perspectives et challenges associés.

2. État de l'art

2.1. Base de données spatio-temporelles

Une base de données spatio-temporelles contient des informations caractérisées par une dimension spatiale et une dimension temporelle. Plus formellement, on peut définir ces bases de données de la manière suivante :

DÉFINITION 1. — *Une base de données spatio-temporelles est un ensemble structuré d'informations défini 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 *instance* matérialisant une zone, un objet ou un événement localisé. Les instances sont liées

par une relation spatiale notée $voisin$ définie par : $voisin(Z_i, Z_j) = vrai$ si Z_i et Z_j vérifient la relation spatiale, $faux$ 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.

Deux types de bases de données spatio-temporelles sont principalement considérées : celles étudiant les trajectoires d'objets qui évoluent dans l'espace et le temps (p. ex. des trajectoires d'oiseaux, d'avions) ; celles étudiant les dynamiques spatiales et temporelles d'événements (p. ex. évolution de l'érosion dans une région ou propagation d'une épidémie dans une ville). Nous allons décrire les données associées à ces deux types de bases et donner des exemples d'interrogations qu'il est possible de formuler à partir de telles données. Pour cela, nous illustrons les concepts sur deux exemples : les déplacements d'avions et les phénomènes climatiques.

2.1.1. Trajectoire d'objets mobiles

Un premier type de bases de données spatio-temporelles consiste à collecter des informations relatives aux déplacements d'objets au cours du temps. Intuitivement, les trajectoires peuvent être vues comme des ensembles de points localisés dans l'espace et le temps (*time-stamped coordinates*). $T =< (t_1, x_1, y_1), ..., (t_n, x_n, y_n) >$ est une trajectoire, i. e. un ensemble de positions (x_i, y_i) de l'objet étudié aux temps t_i. La figure 1 est un exemple de trajectoire qui permet de suivre des objets (ici deux avions) à trois dates. Dans la littérature, on trouve de nombreuses définitions de trajectoires correspondant à des types de déplacement étudié tels que les *flocks* (Gudmundsson, Kreveld, 2006), *moving clusters* (Kalnis *et al.*, 2005), *convoy queries* (Jeung *et al.*, 2008), *closed swarms* (Li *et al.*, 2010), *group patterns* (Y. Wang *et al.*, 2006), *periodic patterns* (Mamoulis *et al.*, 2004)...

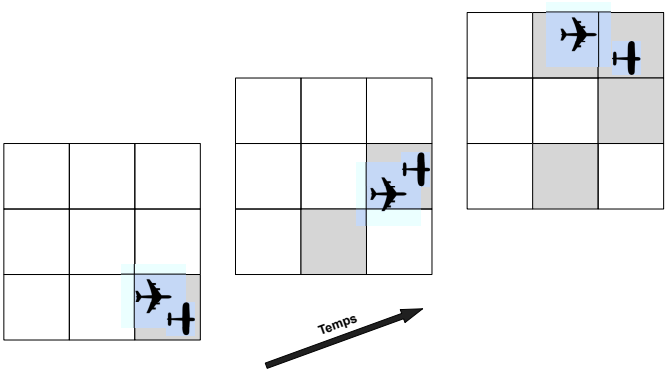


Figure 1. Exemple de trajectoires de véhicules

La dimension temporelle D_T des trajectoires dépend de la manière dont on capture les dates associées aux positions des objets : 1) les positions des objets sont enregistrées à un intervalle de temps régulier, p. ex. toutes les 5s le GPS d'un avion enregistre la position de cet avion (*time-based recording*) ; 2) un enregistrement de la position

se fait à chaque fois que l'objet change de position, p. ex. à chaque fois que l'avion change de route (*change-based recording*) ; 3) un enregistrement se fait quand l'objet se rapproche d'une position, p. ex. à chaque passage de l'avion près d'un aéroport (*location-based recording*) ; 4) l'enregistrement se fait pendant des évènements prédéfinis, p. ex. à chaque appel du pilote à la tour de contrôle (*event-based recording*). Des caractéristiques comme la durée ou la périodicité sont généralement associées à la composante temporelle.

La dimension spatiale D_S des trajectoires dépend de l'information que l'on garde pour localiser l'objet étudié. Celui-ci peut être positionné sur une carte à l'aide de coordonnées géographiques (p. ex. latitude, longitude), polaires ou toutes autres informations (même textuelle). Des caractéristiques comme la direction ou le type de déplacement (en ligne droite, curviligne, circulaire...), les points d'inflexions, peuvent également être associés à la composante spatiale.

L'étude des trajectoires d'objets mobiles est initialement centrée sur le déplacement des objets dans l'espace (autrement dit, sur les dimensions spatiales et temporelles). Par la suite, les dimensions d'analyse D_A ont été enrichies avec des informations statiques sur les objets mobiles eux mêmes (p. ex. le type de l'avion, ou le nombre de passagers...).

Le tableau 1 montre un exemple de base de données d'objets mobiles. Cette base de données trace le déplacement de différents objets sur trois jours consécutifs. Le tableau contient le nom de l'objet et sa position à une date donnée. Nous avons donc $D_T = \{Date\}$, $D_S = \{x_i, y_i\}$ et D_A=$\{Objet\}$. Nous avons choisi de représenter des données correspondant à des temps discrets mais des données estampillées par des temps continus existent également.

Tableau 1. Exemple de base de données d'objets mobiles

$Objet$	$Date$	x_i	y_i
$Objet_1$	t_1	x_1	y_1
$Objet_1$	t_2	x_2	y_2
$Objet_1$	t_3	x_3	y_3
$Objet_2$	t_1	x_4	y_4
$Objet_3$	t_1	x_5	y_5
...	...	...	...

À partir d'une telle base, une requête classique est : *au cours de l'année passée, combien de vols Air France ont été en retard d'au moins 5 minutes au départ de Paris ?* Ce type de requête peut être étendu aux aspects spatiaux dans les applications GIS comme par exemple : *quels sont les avions présents dans l'espace aérien de l'aéroport Charles de Gaule entre 10h et 12h ?* ou *Combien d'avions survolent la région de Toulouse le lundi ?* On peut utiliser ces informations pour déclencher une alerte : *à chaque fois qu'un avion se présente dans l'espace aérien près de l'aéroport de Toulouse* ou établir une requête prédictive : *quels avions vont atteindre Paris dans les 30 minutes ?*

2.1.2. Evolution d'évènements et d'objets localisés

Un deuxième type de base de données spatio-temporelles consiste à stocker des informations sur des évènements (ou des objets) localisés dans l'espace et le temps. On considère ici des zones localisées dans l'espace dans lesquelles se déroulent des événements dont on connaît l'estampille temporelle. $E =< (z_i, t_i, e_i) >$ décrit un évènement e_i se déroulant dans une zone z_i au temps t_i. La figure 2 présente deux exemples de données. La première est une base de données météorologique répertoriant des évènements météo (p. ex. pluie, vent, nuage, soleil) associés à 9 zones sur trois temps consécutifs. La deuxième est une base de données environnementales répertoriant des zones représentées par des polygones et des évènements représentés par des points à trois temps consécutifs. Une des zones de la région étudiée est une rivière dont la surface varie au cours du temps selon les crues.

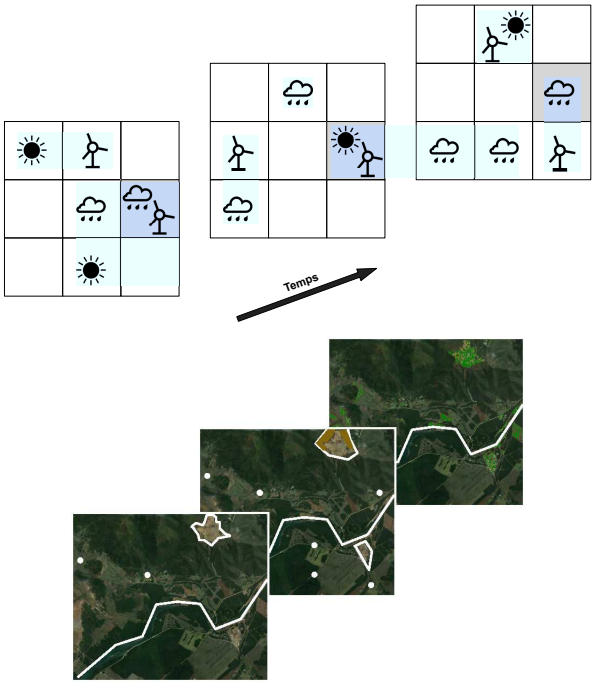


Figure 2. Exemple d'évènements localisés

De même que précédemment, la dimension temporelle D_T d'évènements localisés dépend de la manière dont on capture l'apparition de ces évènements : 1) tous les évènements sont enregistrés à intervalle de temps réguliers dans une zone, p. ex. une balise météorologique enregistre différents paramètres toutes les 5 minutes (*time-based recording*) ; 2) un enregistrement se fait à chaque fois qu'une dimension d'analyse particulière dépasse un seuil, p. ex. à chaque fois que la température dépasse un seuil (*alert-based recording*) ; 3) l'enregistrement se fait par identification du matériel

à des dates prédéfinies, p. ex. l'unité P0221 fait un enregistrement le 04/02/2012 enter 13h00 à 14h10 (*identifier-based recording*).

La dimension spatiale D_S correspond à la définition des zones dans lesquelles se déroulent les évènements. En fonction de la granularité spatiale, la zone z_i peut correspondre à une région (p. ex. une région administrative), à un objet (p. ex. une montagne) ou à un point si l'évènement est très localisé (p. ex. une localisation via GPS). Ces zones peuvent être réparties dans l'espace de manière plus ou moins homogènes. Par exemple, elles peuvent être représentées sous la forme d'une grille aux formes variées (pavage carré, octogonal...) ou sous la forme de polygones ayant des frontières communes (p. ex. les quartiers d'une ville). Dans d'autres cas, les zones sont réparties de manières éparses dans l'espace, avec des possibilités de chevauchement, et ont des formes très variées (polygones, lignes, points). On peut considérer par exemple une découpage administratif (p. ex. une région), qui se superpose à un découpage géographique (p. ex. une montagne, une plaine). Ces zones peuvent être étudiées en fonction de différentes proximités spatiales (p. ex. "à côté de", "près de", "en amont de"...).

Les dimensions d'analyse D_A correspondent aux différents types d'évènements et d'objets étudiés. Chaque dimension est associée ensuite à un domaine de valeurs représentant les différentes propriétés/caractéristiques des évènements et objets étudiés (p. ex. pluie faible/forte).

Le tableau 2 illustre un exemple de base répertoriant des évènements météorologiques. Cette base de données associe des évènements météo 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 les relations de proximité suivantes ; la zone Z_1 est située à côté des zones Z_2 et Z_3. Ces dernières ne sont pas voisines. Nous avons donc $D_T = \{Date\}$, $D_S = \{Ville\}$ 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 $dom(Rafales) = \{55, 75\}$. Il est possible également d'associer des informations statiques aux zones (p. ex. une surface, un relief...).

À partir d'une telle base, il est possible de formuler des requêtes telles que : *quelle zone a été la plus pluvieuse au cours du temps ?* Des requêtes prédictives peuvent être posées telles que : *dans quelle zone peut-on prévoir une augmentation du vent ?* ou *S'il pleut au nord de la ville à midi, quelles seront les conséquences dans l'après midi au sud ?*

Dans la suite de cet état de l'art, nous nous intéressons aux différentes méthodes permettant d'extraire de l'information à partir de ces deux types de bases. Nous commençons par les méthodes dédiées au suivi des trajectoires, puis nous continuons avec

Tableau 2. Exemple de base de données d'évènements localisés

Ville	Date	Température	Précipitation	Vent	Rafales
Z_1	$T_1 = 22/12/10$	T_m	P_m	V_m	-
Z_1	$T_2 = 23/12/10$	T_m	P_m	V_l	-
Z_1	$T_3 = 24/12/10$	T_l	P_m	V_m	55
Z_2	$T_1 = 22/12/10$	T_m	P_m	V_m	-
Z_2	$T_2 = 23/12/10$	T_l	P_m	V_l	-
Z_2	$T_3 = 24/12/10$	T_l	P_l	V_m	-
Z_3	$T_1 = 22/12/10$	T_l	P_m	V_s	75
Z_3	$T_2 = 23/12/10$	T_m	P_s	V_l	-
Z_3	$T_3 = 24/12/10$	T_l	P_s	V_s	55
…	…	…	…	…	…

deux approches dédiées aux bases d'évènements localisés : les motifs spatiaux et les motifs spatio-temporels.

2.2. Trajectoires pour le suivi d'objets mobiles

L'émergence des nouvelles technologies mobiles a entraîné la collecte de grandes quantités de données spatio-temporelles, dédiée à la localisation d'objets mobiles dans l'espace et le temps comme nous venons de le décrire dans la section 2.1.1. Ces nouvelles bases permettent d'entrevoir de nouvelles applications. Par exemple, le projet GeoPKDD (Giannotti, Pedreschi, 2008) a étudié l'aménagement du plan de circulation de grandes agglomérations en fonction des déplacements des véhicules. On trouve d'autres domaines d'applications : la géographie socio-économique, le sport (p. ex. les joueurs de football), l'analyse et le contrôle de la pêche, les prévisions météorologiques (p. ex. des ouragans). Dans la plupart de ces applications, le nombre de trajectoires est important. L'un des objectifs des méthodes d'analyse de ces trajectoires est de trouver les plus pertinentes selon l'objectif applicatif (p. ex. les plus fréquentes, les plus inattendues, les périodiques...). Face à cette problématique, plusieurs approches ont été proposées dans la littérature. Nous nous focalisons dans la suite de cette section aux méthodes de fouille de données.

Dans (Mamoulis *et al.*, 2004 ; Cao *et al.*, 2005 ; 2007), les auteurs se sont intéressés à l'extraction de motifs périodiques. Les objets étudiés, des bus, ont la particularité de suivre approximativement la même route à intervalles de temps réguliers. Dans un premier temps, cette approche consiste à résumer l'ensemble des trajectoires d'un même objet par une seule séquence de segments. Les segments de trajectoires similaires sont regroupés en utilisant une fonction de similarité qui tient compte de la proximité spatiale, basée sur l'angle et la longueur spatiale des segments. Cette approche donne une meilleure abstraction des trajectoires et diminue la taille des données pour l'extraction. Contrairement à d'autres travaux, ces motifs intègrent une notion floue au niveau des localisations, ce qui permet d'extraire un motif même s'il ne se répète pas exactement au même endroit. Les auteurs ont utilisé un algorithme par niveau dérivé d'$Apriori$, et l'ont optimisé grâce à l'utilisation d'une nouvelle structure de données (substring tree). Ce travail a été validé sur une base de données de trajectoires de bus.

Si l'on reprend l'exemple des avions, chaque séquence correspond aux déplacements d'un avion au cours d'une semaine.

Dans (Fisher *et al.*, 2005), les motifs étudiés sont des groupes d'objets partageant un type de mouvements (direction, vitesse) à une date donnée dans une certaine région de l'espace. Cinq types de motifs de trajectoires basés sur le mouvement, la direction et la localisation sont proposés (convergence, rencontre, troupeau, leadership et récurrence). Les travaux présentés dans (Gudmundsson *et al.*, 2004) permettent de détecter les 4 premiers types de motifs définis dans (Fisher *et al.*, 2005) en utilisant des algorithmes de calcul approximatif. Un exemple de motif que nous pourrons extraire est : les vols AF3234, AF7681 et KL8089 ont été en concurrence dans une période de 30 minutes quand ils survolaient Bordeaux avec une orientation constante de 33°NE et une vitesse de 830 Km/h.

Nanni et Pedreschi (2006) présentent une adaptation d'un algorithme de clustering basé sur la densité pour les trajectoires d'objets en mouvement. Ils s'appuient sur une notion de distance entre trajectoires. Ils mettent l'accent sur la dimension temporelle - essentiellement en élargissant l'espace de recherche des groupes intéressants en tenant compte des restrictions des trajectoires sources sur des sous-intervalles de temps. L'algorithme proposé vise à chercher les intervalles de temps les plus significatifs qui permettent d'isoler les groupes de qualité supérieure. Grâce à cette contribution, nous pouvons extraire l'information : le trafic aérien a été très agité à l'aéroport Charles de Gaulle le 22 mai de 9h34 à 9h37 à cause d'un grand nombre de vols au départ initialement annoncés cumulés avec ceux ayant du retard.

Giannotti *et al.* (2007) proposent une extension du paradigme d'extraction de motifs séquentiels à l'analyse des trajectoires. Ils introduisent les motifs de trajectoires comme des descriptions concises de comportements fréquents, en termes d'espace (les régions de l'espace visitées lors des déplacements) et de temps (la durée des déplacements). Ce travail est davantage axé sur des concepts de niveau supérieur (au lieu de découvrir un motif impliquant un endroit spatial précis, une localisation générale est trouvée). Ces localisations générales sont appelées régions d'intérêt (*Regions-of-Interest* ou *RoI*). Les motifs fréquents de déplacement entre ces régions sont découverts par la suite. Par exemple, pour se déplacer de Montpellier vers l'Asie, la plupart des vols (plus du 70 %) passent d'abord par Paris et après par Amsterdam. Ces deux aéroports sont des points d'intérêt.

Nous détaillons dans les deux sections suivantes, les travaux de la littérature relatifs à l'étude d'évènements localisés tels que définis dans la section 2.1.2, que nous illustrons sur l'exemple des phénomènes climatiques. Nous nous intéressons à deux familles de motifs : les motifs spatiaux et les motifs spatio-temporels.

2.3. Motifs spatiaux et spatio-temporels pour le suivi d'évènements localisés

L'extraction de motifs spatiaux et spatio-temporels a été largement étudiée ces dernières années dans des données géographiques et des SIG. Il y a deux familles

d'approches : les approches basées sur les colocalisations (Shekhar, Huang, 2001) (*colocations* en anglais) qui identifient des évènements fréquemment proches et les motifs spatio-temporels qui identifient l'évolution d'évènements dans l'espace et le temps.

Les colocalisations se focalisent sur les objets et leurs relations spatiales. Elles ont été proposées dans (Shekhar, Huang, 2001) avant d'être étendues dans (Huang *et al.*, 2004 ; Yoo, Shekhar, 2006 ; Celik *et al.*, 2007 ; 2008 ; Lin, Li, 2009 ; L. Wang *et al.*, 2009). L'objectif est de trouver tous les sous-ensembles de propriétés (ou types d'évènements) fréquemment associés à des objets spatiaux voisins. Le motif $\{nuage, pluie, montagne\}$ dans la figure 3 est un exemple de colocalisation qui traduit le fait que les nuages s'accrochent au relief ce qui entraîne de la pluie. Les propriétés $nuage, pluie,$ et $montagne$ sont souvent corrélées spatialement. Une mesure d'intérêt anti-monotone a été introduite, appelée l'indice de participation, pour filtrer les colocalisations les plus importantes. Cette mesure traduit globalement la fréquence de participation des caractéristiques à une colocalisation par le biais de leurs instances. Un algorithme par niveaux extrait les solutions.

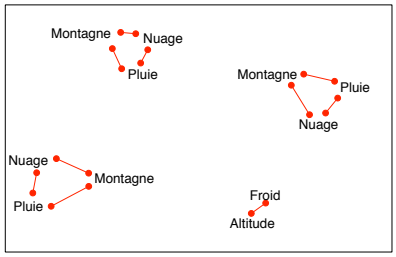


Figure 3. Exemple de colocalisation

Comme pour les algorithmes d'extraction d'itemsets, Bogorny *et al.* (2006) se sont intéressés à l'intégration des contraintes dans l'extraction de motifs spatiaux. Toutefois, leur travail s'appuie sur une approche orientée transactions qui ne prend en compte que partiellement les relations spatiales et fait un prétraitement des données pour assimiler des contraintes expertes. L'idée de base est de considérer une caractéristique de référence, d'énumérer d'abord les voisinages pour construire un ensemble de transactions autour des instances de la caractéristique spatiale de référence. Les auteurs appliquent ensuite une méthode d'extraction d'itemsets, avec des contraintes expertes (par exemple, éliminer tous ceux qui sont connus par l'expert), dans la base de données transactionnelles obtenue. Par exemple, la caractéristique de référence est "station d'essence", une base de données transactionnelles générée considérant les caractéristiques de toutes les instances proches des instances de la référence, l'itemset fréquent "Station d'essence, Cours d'eau, Pollution" est extrait et considéré comme une colocalisation pertinente à la caractéristique de référence "station d'essence".

Si ces motifs sont très intéressants en permettant de révéler des associations entre des évènements et des lieux, il ne permettent pas de matérialiser l'évolution et le

déplacement de ces évènements. Par exemple, dans le cas de la météo, on ne peut capturer des informations sur des évolutions comme la transformation d'un nuage spécifique en orage près de courants d'air ascendants existant près des montagnes. Au contraire, l'objectif des motifs spatio-temporels, est d'étudier l'évolution et les interactions globales, dans l'espace et dans le temps, d'ensembles d'évènements.

Dans cette catégorie de méthodes, Celik *et al.* (2006 ; 2008) ont généralisé le concept de colocalisations à des données spatio-temporelles. Les colocalisations spatio-temporelles représentent des ensembles de propriétés associées à des objets voisins dans l'espace et dans le temps. Plus précisément, ils représentent des instances spatialement proches pendant une fraction significative de temps. Une mesure d'intérêt monotone combinant prévalence spatiale et prévalence temporelle permet d'intégrer conjointement cette proximité spatiale et temporelle. Dans la figure 4, les motifs $\{temperatureSous0, pic\}$ et $\{nuage, pluie, grele\}$ sont des colocalisations spatio-temporelles (les traits en pointillés représentent la relation de voisinage). Pour extraire ces motifs, les auteurs ont proposés une stratégie générer-tester et un parcours par niveaux de type $Apriori$.

Ce concept a aussi été étudié dans (Qian *et al.*, 2009). Les auteurs se sont intéressés aux "trajectoires" de colocalisations spatiales, appelées SPCOZ (*Spread patterns of spatio-temporal co-occurrences over zones*). Un élément de propagation est une colocalisation "fréquente" associée à une fenêtre temporelle. Dans la figure 4, la colocalisation fréquente $\{altitude, froid\}$ associée à l'intervalle [t0,t1] est un élément de propagation. Combinés deux à deux, ces éléments constituent des arbres représentant la propagation (*SP-Tree* ou *Spread Pattern Tree*). L'algorithme d'extraction commence par rechercher toutes les colocalisations fréquentes de taille 2 (par une méthode de type $Apriori$), les utilise pour construire les éléments de propagation pour tous les temps, et génère les *SP-Tree* correspondants.

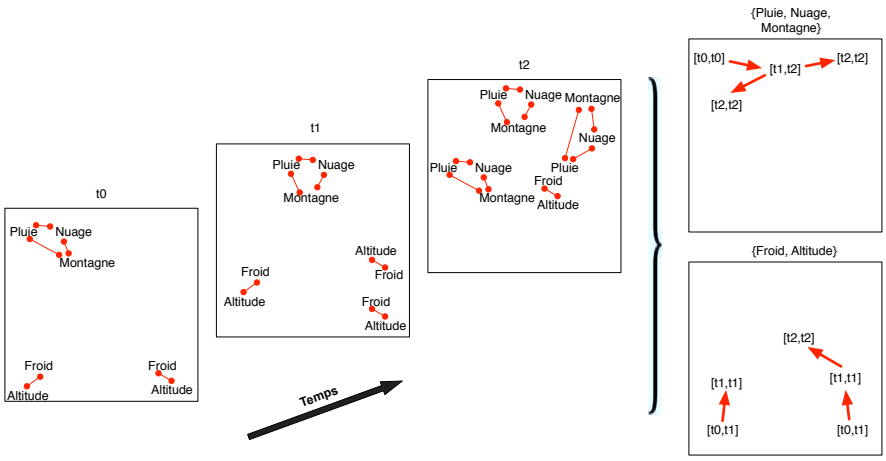


Figure 4. Exemple de colocalisations spatio-temporelles et de SPCOZ

Dans (Yang *et al.*, 2005), les auteurs proposent un "framework" pour l'extraction de motifs spatiaux. Ils ont validé leur approche sur un jeu de données étudiant l'évolution de molécules et de vortex. Les motifs étudiés, appelés SOAP (*Spatial Object Association Pattern*), peuvent être représentés sous la forme de graphes où, chaque nœud correspond à une propriété et chaque arête représente une relation de voisinage. Contrairement aux colocalisations, les auteurs considèrent des objets géométriques plutôt que des points pour les calculs de voisinage et extraient trois autres types de configuration : étoile, séquence et $minLink$. Ce dernier type correspond à des SOAP plus généraux où seul le nombre minimum d'arêtes ($minLink$) associées à chaque nœud est fixé. Finalement, les auteurs montrent aussi comment utiliser (en post-traitement) les SOAP fréquents pour visualiser l'évolution d'un même ensemble de propriétés F. Pour cela, ils définissent la notion d'épisodes comme un ensemble d'instances associées à un intervalle de temps où le motif est apparu puis disparu. La séquence ainsi générée permet de visualiser l'évolution spatiale de l'ensemble des propriétés étudiées. Toutefois, ce post-traitement ne permet pas de prendre en compte les évolutions de forme des objets étudiés, ainsi que les éventuelles relations de cause à effet. De plus, le nombre d'épisodes peut être important pour un même ensemble de propriétés, ce qui rend très difficile l'analyse de la séquence.

Les séquences et plus généralement les graphes, ont été souvent utilisés et étendus au spatio-temporel pour représenter la *propagation* de phénomènes dans l'espace et dans le temps (Tsoukatos, Gunopulos, 2001 ; J. Wang *et al.*, 2004 ; 2005 ; Huang *et al.*, 2008 ; Mabit *et al.*, 2011 ; Selmaoui-Folcher, Flouvat, 2011 ; Alatrista-Salas *et al.*, 2011 ; 2012).

Dans (J. Wang *et al.*, 2004), les auteurs se focalisent sur la propagation spatio-temporelle d'événements dans des fenêtres temporelles prédéfinies. Ils découpent la dimension temporelle en fenêtres, divisent l'espace sous la forme d'une grille. Ils définissent le concept de flow pattern comme une séquence d'ensembles d'évènements de la forme $< E_1 \rightarrow ... \rightarrow E_k >$ où E_i est un ensemble d'événements de la forme $e(localisation)$, avec e un type d'évènements (p. ex. pluie, vent). Chaque ensemble d'évènements est composé d'évènements spatialement voisins apparaissant au même temps. Deux ensembles d'évènements E_p et E_q sont consécutifs dans la séquence, si leurs évènements appartiennent à la même fenêtre temporelle, s'ils sont tous voisins et qu'ils apparaissent à deux temps consécutifs. À titre d'exemple, le flow pattern $< \{nuage(0,0)\} \rightarrow \{pluie(1,0)\} >$ signifie que des nuages apparus en position (0,0) ont précédé légèrement (le laps de temps dépend de la fenêtre temporelle étudiée) l'apparition de pluie en position (1,0) (une position voisine). Pour extraire ces motifs, les auteurs appliquent une stratégie par niveaux pour trouver les séquences de taille 1 et 2, puis utilisent les motifs fréquents trouvés comme point de départ à un parcours en profondeur de l'espace de recherche.

Dans un deuxième temps, (J. Wang *et al.*, 2005) étendent cette notion et définissent les motifs spatio-temporels généralisés (*Generalized spatio-temporal pattern*) comme des séquences de *relative eventsets*. Un *relative eventset* est un ensemble d'évènements dont la localisation est remplacée par un positionnement relatif à une locali-

sation de référence. Un motif spatio-temporel généralisé est fréquent s'il a au moins t-minsup (support temporel) occurrences dans le temps et qu'il a au moins s-minsup (support spatial) occurrences dans l'espace (les localisations peuvent être différentes mais la localisation relative doit être identique). Pour extraire ces motifs, les auteurs proposent un nouvel algorithme appelé GenSTMiner qui utilise une approche dérivée de Prefispan (Pei *et al.*, 2004). Nous pouvons extraire, par exemple, la séquence traduisant le fait que "des températures élevées et de l'humidité apparaissent fréquemment dans une zone après l'apparition de forte pluie dans une zone voisine, pendant une semaine".

(Huang *et al.*, 2008), se sont concentrés sur le problème d'extraction de séquences de propriétés représentant la propagation de certains types d'évènements. Ces séquences sont de la forme $< f_1 \rightarrow f_2 \rightarrow ... \rightarrow f_k >$, où f_i est un type d'évènements. Cette approche permet donc d'étudier la propagation des évènements pris individuellement (sans prendre en compte leur environnement). Ce modèle considère deux évènements comme consécutifs s'ils sont spatialement proches (distance euclidienne inférieure à un seuil donné) et apparaissent dans la même fenêtre temporelle. Les auteurs ont également étudié d'autres relations de voisinage dépendant du temps. Ces relations permettent de représenter un rétrécissement de la zone d'influence d'un évènement (son voisinage) au cours du temps. Les auteurs proposent aussi une nouvelle mesure d'intérêt pour ces séquences qui reflètent plus un lien de cause à effet entre les évènements. Cette mesure n'étant pas anti-monotone, les auteurs proposent donc un nouvel algorithme $Slicing-STS-Miner$, pour extraire ces séquences, basé sur un traitement incrémental des différentes fenêtres temporelles et une extension des séquences à chaque étape. A défaut de l'anti-monotonie, ils exploitent une autre propriété du *sequence index* : si une séquence est intéressante, toutes les sous-séquences ayant le même préfixe sont intéressantes.

Dans la suite de cet article, nous allons présenter deux méthodes de fouille de données en nous focalisant uniquement sur l'étude des objets spatialisés. La première méthode porte sur les colocalisations et la deuxième sur les motifs spatio-séquentiels. La première méthode a été appliquée à l'étude de l'érosion des sols et la deuxième à l'étude des épidémies de dengue.

3. Colocalisations sous-contraintes pour caractériser l'érosion

Nous présentons, dans cette section, une approche basée sur l'extraction de colocalisations satisfaisant des contraintes thématiques définies par les experts. Cette méthode a notamment été appliquée à la caractérisation de l'érosion. En plus de la contrainte liée à la probabilité d'apparition, l'algorithme d'extraction vérifie des contraintes basées sur la connaissance des experts. Ces contraintes ont l'avantage d'améliorer la pertinence des motifs tout en réduisant l'espace de recherche et en améliorant les performances.

3.1. Les données

Dans cet exemple, nous sommes dans le cas particulier où l'on souhaite étudier les corrélations spatiales d'objets ou d'événements. La dimension temporelle n'étant pas présente, la base de données utilisée peut être considérée comme le couple (D_S, D_A) où D_S représente la dimension spatiale et D_A l'ensemble des dimensions qui décrivent les attributs que l'on appelle ici caractéristiques. Chaque D_{A_i} représente le domaine d'un thème ou d'une couche par exemple le thème végétation dont le domaine peut être végétation dense, végétation éparse, etc.

La méthode est illustrée sur un exemple de données réelles liées à l'érosion des sols. Ces données ont été préparées par des experts géologues. Elles concernent un bassin versant montagneux calédonien de $9\ km^2$ présentant de l'érosion naturelle et de l'érosion liée à des activités minières. Nous disposons d'informations décrivant des objets qui sont des zones géographiques classées selon 3 thèmes : type d'érosion, nature du sol et le type de végétation. Les colocalisations permettent de décrire des associations fortes (très fréquentes) entre des caractéristiques avec un taux de participation quantifié par la mesure "indice de participation" définie dans le paragraphe ci-dessous. Ce résultat permet à l'expert de savoir dans quelles zones approximativement et à côté de quels types d'objets il y a un risque d'un type d'érosion. Un exemple de motif recherché montre une association forte entre zone "piste sensible", zone "minière", zone "érosion en rivière" et "végétation éparse".

3.2. Les colocalisations sous-contraintes

3.2.1. Le concept de colocalisation

Considérons un ensemble de caractéristiques (ou attributs) catégorielles $\mathcal{D}_\mathcal{A} = \{D_{A_1}, D_{A_2}, ..., D_{A_k}\}$. Chaque dimension D_{A_i} représente un thème (par exemple végétation) dont le domaine $Dom(A_i) = \{I_{i1}, ..., I_{ip}\}$ est un ensemble d'items ou caractéristiques (p. ex. végétation dense, végétation éparse...). Considérons également un ensemble d'objets géolocalisés $\mathcal{D}_\mathcal{S} = \{z_1, z_2, ..., z_n\}$ associés à ces caractéristiques (p. ex. des régions). Chaque objet spatial est un vecteur <*id de l'objet, attributs catégoriels, localisation*>.

DÉFINITION 2. — *Une* ***colocalisation*** C *est un sous-ensemble de caractéristiques de* $\mathcal{D}_\mathcal{A}$ *tel qu'il existe des ensembles d'objets voisins ayant ces mêmes caractéristiques.*

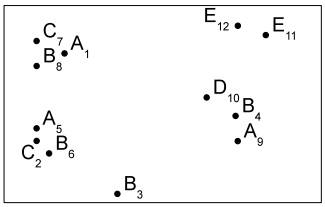


Figure 5. Représentation graphique des objets spatiaux et de leur caractéristique

Dans l'exemple de la figure 5, $\{A, B, D\}$ est une colocalisation car ces caractéristiques sont associées aux objets voisins notés D_{10}, B_4 et A_9 sur la figure.

Une **instance** d'une colocalisation C, par rapport à une relation de voisinage $\mathcal{R}$ fixée, est un ensemble d'objets de $\mathcal{D}_\mathcal{S}$ ayant pour caractéristiques celles de C, la même taille que C, et respectant deux à deux la relation spatiale $\mathcal{R}$. Sur la figure 5, $\{A_1, B_8, C_7\}$ est une instance de la colocalisation $\{A, B, C\}$ (A_1, B_8, et C_7 voisins d'après $\mathcal{R}$).

La **tableau d'instances** d'une colocalisation C, notée $TI_\mathcal{R}(\mathcal{D}_\mathcal{S}, C)$ est l'ensemble des instances de C. Elle correspond à tous les ensembles d'objets de $\mathcal{D}_\mathcal{S}$ vérifiant la colocalisation C par rapport à la relation de voisinage $\mathcal{R}$. Dans la figure 5, le tableau d'instances de $\{A, B, C\}$ est $TI_\mathcal{R}(\mathcal{D}_\mathcal{S}, ABC) = \{ \{A_1, B_8, C_7\}, \{A_5, B_6, C_2\} \}$ et le tableau d'instances de $\{B, D\}$ est $TI_\mathcal{R}(\mathcal{D}_\mathcal{S}, BD) = \{ \{B_4, D_{10}\} \}$

Shekhar et Huang (2001) ; Huang *et al.* (2004) ont introduit la notion d'**indice de participation** comme mesure de fréquence, notée pi, représentant la probabilité minimale d'avoir un objet ayant une caractéristique de la colocalisation C parmi l'ensemble des objets ayant cette même caractéristique.

$$pi_\mathcal{R}(\mathcal{D}_\mathcal{S}, C) = \min_{\forall I \in C} (pr_\mathcal{R}(\mathcal{D}_\mathcal{S}, C, I))$$

où :

$$pr_\mathcal{R}(\mathcal{D}_\mathcal{S}, C, I) = \frac{|\{z \in Z \mid Z \in TI_\mathcal{R}(\mathcal{D}_\mathcal{S}, C) \text{ et } z \text{ véfifie la caractéristique } I\}|}{|TI_\mathcal{R}(\mathcal{D}_\mathcal{S}, I)|}$$

3.2.2. Intégration des contraintes du domaine et extraction

Nous nous sommes appuyés sur le cadre théorique de Mannila et Toivonen (1997) afin de pouvoir intégrer de nouvelles contraintes expertes au cœur de l'algorithme (évitant ainsi des pré ou post traitements coûteux en performances). Le cadre théorique de l'extraction des colocalisation a ainsi été généralisé à tout prédicat anti-monotone, sans pour autant remettre en question les aspects algorithmiques (Flouvat *et al.*, 2010). La connaissance des experts est intégrée sous la forme d'une conjonction de contraintes anti-monotones. Cette proposition permet d'avoir une information plus fine en sortie de l'analyse en filtrant des colocalisations non pertinentes au cours du processus d'extraction.

Dans le contexte de l'étude de l'érosion, les contraintes vont représenter des relations déjà connues par les experts ou non intéressantes, et qui devront être éliminées des résultats. Ces contraintes peuvent être perçues comme des règles d'exclusion. L'intérêt de cette approche est d'obtenir une information plus pertinente en sortie de l'analyse, tout en améliorant l'efficacité des algorithmes.

L'une des particularités des SIG est la représentation de l'information en couches ou thèmes, chaque couche contenant un ensemble d'objets géographiques, eux-mêmes associés à un ensemble de caractéristiques. L'intégration des contraintes expertes prend en compte ce découpage de l'information en permettant de définir des

contraintes sur les caractéristiques, les thèmes [2] et les objets. Nous considérons plus particulièrement deux types de contraintes :

– les contraintes sur les caractéristiques et les thèmes ;

– les contraintes spatiales sur les objets.

Le premier type de contraintes va permettre à l'expert de spécifier finement des colocalisations à ne pas étudier. Par exemple, l'expert n'est pas intéressé par les relations entre les caractéristiques *érosion de versant* et *harzburgites*. Il peut aussi utiliser ces contraintes pour focaliser son étude sur certaines problématiques (en excluant toutes les caractéristiques/thèmes qui n'y sont pas liés).

Un cadre théorique et un formalisme on été introduits dans (Flouvat *et al.*, 2010) afin de représenter les contraintes, et de les intégrer sous-forme de prédicats dans l'algorithme d'extraction. Les colocalisations contenant un ensemble de caractéristiques et/ou abordant un ensemble de thèmes peuvent ainsi être éliminées de l'analyse. Il devient, par exemple, possible d'exclure les colocalisations étudiant la caractéristique *sol non nu* conjointement avec le thème *Végétation*.

Les résultats obtenus par ce type de contraintes sont fondamentalement différents de ce que l'on aurait eu en faisant un prétraitement classique tel que "supprimer des données les caractéristiques à exclure". Par exemple, supposons que l'on ignore lors de l'analyse les données liées à *érosion de versant* et à *harzburgites*. Ce pré-traitement permettrait d'éviter l'étude des relations entre ces deux caractéristiques, mais empêcherait aussi d'étudier les relations entre *érosion de versant* et la végétation. A l'opposé, nos contraintes permettent d'étudier toutes ces relations et évitent uniquement l'étude des motifs concernant le lien entre *érosion de versant* et *harzburgites*.

Le deuxième type de contraintes va permettre à l'expert d'exclure de l'analyse des objets géographiques en fonction de critères spatiaux et de leurs caractéristiques (et/ou thèmes). Une contrainte classique consiste à ne pas étudier les objets situés dans certaines zones géographiques, tout en précisant éventuellement les caractéristiques ou thèmes des objets à considérer. Un exemple de ce type de contraintes serait "ne pas étudier les objets caractérisés par *sol non nu* et *Végétation*, et situés dans une zone rectangulaire ayant pour coordonnées (100,200,400,600)".

Contrairement aux contraintes "thématiques" sur les colocalisations, les contraintes spatiales sur les objets n'apparaissent pas directement dans le prédicat utilisé pour l'extraction des colocalisations intéressantes. Par contre, elles influencent le calcul de l'indice de participation en diminuant le nombre d'instances des colocalisations étudiées. La définition des tableaux d'instances se trouve donc modifiée.

De même que les contraintes sur les caractéristiques et les thèmes, ces contraintes spatiales sur les objets permettent des traitements beaucoup plus "fins" que simplement enlever des données les objets situés dans certaines zones. En effet, elles permettent d'ignorer très localement certains objets à certaines étapes de l'extraction.

2. Thème : ensemble d'objets du même type. Par exemple thème végétation.

Ces contraintes permettent, par exemple, d'ignorer partiellement des zones en raison de problèmes d'acquisition pour certaines caractéristiques, ou d'exclure de l'analyse des relations dans certaines régions lorsque celles-ci sont déjà connues.

3.3. Résultats expérimentaux

3.3.1. Protocole expérimental

Nous avons intégré ces propositions dans un prototype de découverte de motifs s'appuyant sur l'outil iZi développé par Flouvat *et al.* (2009). Cet outil a été complété par un module permettant d'accéder à une base de données géographiques $PostGis$. Ce prototype intègre aussi l'approche de visualisation des colocalisations présentées dans (Desmier *et al.*, 2011). Grâce à cette approche, les experts peuvent visualiser sur une carte les colocalisations extraites (alors qu'habituellement elles sont présentées sous la forme d'un rapport textuel). Ces colocalisations sont géo-localisées sur la carte en fonction de la localisation de leurs instances. Pour une colocalisation donnée, l'expert peut ainsi connaître où et comment sont localisés les évènements qu'elle représente.

Les expérimentations ont été réalisées sur un bassin versant montagneux d'environ 9 km^2. Il présente des manifestations de l'érosion naturelle ainsi que des stigmates liés à l'activité minière. Ce jeu de données est constitué de 7 700 objets associés à des caractéristiques regroupées en trois couches. Les trois couches thématiques considérées comme importantes pour les experts (information spatiale continue) sont :

– La couche "Etat du sol" indique si le sol est soumis à des processus liés à l'érosion à l'échelle du 1/50 000. Elle a été élaborée à partir d'une carte de sols nus établie par télédétection. Une information sur le type de phénomène actif dans chaque zone est disponible, elle a été obtenue par application de règles SIG simulant l'interprétation d'un expert (Rouet *et al.*, 2009). L'état du sol peut être "sol non nu", "sol nu sur piste", "érosion de versant", "érosion sur mine", "érosion en rivière" ou "zone sédimentaire active" ;

– La couche "Nature du terrain" contient une information sur la lithologie. Elle est issue du SIG géologique à l'échelle du 1/50 000 (DIMENC/SGNC, BRGM, 2005). C'est une donnée vectorielle de type polygone pour lesquels 13 types différents de nature lithologique sont distingués ;

– La couche "Végétation" décrit la répartition spatiale au 1/50 000 des différents types de systèmes végétaux présents sur la zone d'étude. Elle a été obtenue par télédétection (DTSI/SGT, 2008).

Les objets d'étude sont donc des surfaces, i. e. des polygones, associés chacune à une caractéristique d'une des trois couches. Ce jeu de données a été étudié par rapport à plusieurs seuils minimum d'indice de participation ainsi que différentes relations de voisinage. Afin de simplifier l'étude de ces expérimentations, nous présentons dans un premier temps les résultats et les performances obtenus à partir d'une relation de voisinage liée à la distance euclidienne. Plus précisément, deux surfaces sont considérées

comme voisines lorsque la distance entre leurs centroïdes est inférieure à un certain seuil. Puis, dans un second temps, nous présentons les conclusions de l'expert à partir de cette relation de voisinage ainsi qu'une autre liée à l'intersection entre surfaces.

3.3.2. Résultats quantitatifs et performances

Le tableau 3 présente le nombre de colocalisations extraites en utilisant des contraintes définies par un expert géologue. Trois contraintes ont été étudiées :

– exclure les colocalisations ayant la caractéristique "Maquis ligno-herbacé" ($Constraint\ 1$) ;

– exclure les colocalisations ayant la caractéristique "Maquis ligno-herbacé" et une caractéristique du thème "Erosion" ($Constraint\ 2$) ;

– exclure les colocalisations ayant la caractéristique "Maquis ligno-herbacé" et une caractéristique du thème "Erosion", et ignorer tous les objets d'une zone donnée ($Constraint\ 3$).

Comme le montre cet exemple, le nombre de motifs découverts peut fortement diminuer en fonction des contraintes (jusqu'à 32 %) améliorant ainsi les performances, l'interprétation et la pertinence des résultats pour les experts.

Tableau 3. Nombre de colocalisations en fonction des contraintes expertes

	$Seuil$	Sans contrainte	Maquis ligno-herbacé ET Erosion	(Maquis ligno-herbacé ET Erosion) OU Zone
$Distance\ 100$	0.1	66	54	37
	0.3	12	11	7
	0.5	4	4	2
$Distance\ 200$	0.1	268	186	124
	0.3	72	59	24
	0.5	22	22	22
$Distance\ 300$	0.1	707	430	246
	0.3	168	123	58
	0.5	57	57	57

La figure 6 confirme le gain en performances. Cette figure compare le temps d'exécution de notre algorithme d'extraction de colocalisations sous contraintes avec l'algorithme d'extraction de colocalisations classique proposé dans (Huang *et al.*, 2004), pour les seuils et distances étudiés précédemment. Les courbes notées $V0$ correspondent aux temps d'exécution de l'algorithme classique de (Huang *et al.*, 2004) (sans contraintes donc). Les courbes notées $Constraint$ correspondent au temps d'exécution de notre algorithme associé aux différentes contraintes. Comme le montrent ces schémas, l'extraction est plus rapide avec les contraintes malgré le surcoût pour les vérifier. Cette accélération est liée à la diminution du nombre colocalisations et d'objets étudiés (cf. tableau 3). Le nombre de colocalisations et d'objets traités avec ces contraintes sont proches ce qui explique la faible différence entre les trois exécutions sous contraintes.

3.3.3. Résultats qualitatifs et retours des experts

Les trois couches ont été analysées par l'expert de plusieurs manières :

– Traitement A : l'étude a dans un premier temps été focalisée sur les centroïdes des surfaces. Il s'agissait d'évaluer l'apport de l'analyse des distances entre centroïdes pour l'expert. Plusieurs distances ont été testées (50, 100 et 200 m) avec différents seuils d'indice de participation.

– Traitement B : les intersections entre les surfaces des trois thèmes ont ensuite été analysées. Les trois types d'intersections (point, limite ou surface) sont considérés indifféremment. Les résultats ont été étudiés à différents seuils minimum d'indice de participation.

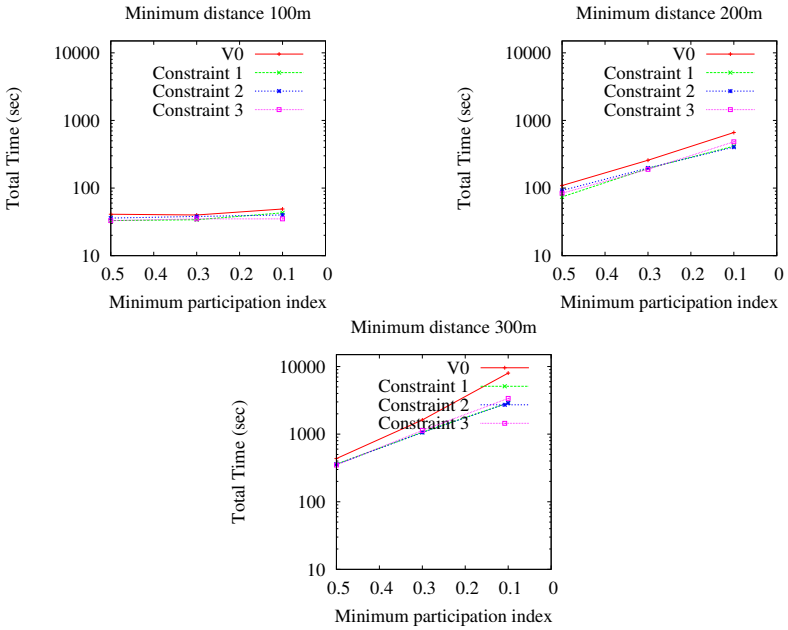


Figure 6. Performances de l'algorithme d'extraction de colocalisations sous contraintes en fonction de différents seuils minimaux d'indice de participation

Les résultats du traitement A montrent des colocalisations à 50 m, à 100 m et à 200 m. Les colocalisations intra-thèmes sont essentiellement liées au thème "Etat du sol", avec un ratio de colocalisation dépassant fréquemment 0,4. Les plus significatives mettent en avant la proximité fréquente des mines à ciel ouvert et des zones d'érosion de versant (0,5 à 50 m, 0,7 à 100 m et 0,9 à 200 m) ou encore l'association de pistes sensibles, mine, érosion de versant et érosion en rivière (0,57 à 200 m). De nombreuses colocalisations au sein de la couche "Végétation" apparaissent aussi à partir de 100 m de distance entre centroïdes, elles sont de taille 2 ou 3. Elles permettent de relever des associations de systèmes végétaux fréquentes (végétation arbustive et forêt sur substrat volcano-sédimentaire avec 0,5 à 100 m, 0,6 à 200 m). Aucune colocalisation

intra-thème n'a été identifiée pour la nature du terrain. Cet écart entre les résultats s'explique aisément par le contenu même des thèmes. L'état du sol et la végétation sont issus de l'analyse de la même scène satellitale avec un niveau de détail plus fin que pour la nature du sol. Les seuils de distance choisis ne permettent pas de faire ressortir les colocalisations dans cette dernière couche qui peut présenter des polygones de grande superficie, soit une distance entre centroïdes au-delà des seuils choisis. Les colocalisations inter-thèmes ne sont mises en évidence par le traitement A qu'à partir de 200 m. Les plus éloquentes sont les associations entre pistes sensibles, zones minières, érosion en rivière et végétation éparse (0,37 à 200 m) et entre mines, érosion de versant, maquis ligno-herbacé et pistes sensibles (0,32 à 200 m) ou érosion en rivière (0,35 à 200 m). Elles soulignent très nettement la dégradation du milieu aux alentours des zones où les sols ont été décapés par l'homme.

Les résultats du traitement B apportent quant à eux un autre type d'informations. Les colocalisations par intersection peuvent être de taille 2 ou 3. Cette fois, des colocalisations intra-thèmes sont mises en évidence pour les trois thèmes, avec 7 sur la nature du sol, 9 sur l'état du sol et 10 sur la végétation. Par exemple, les résultats montrent que les zones minières et à érosion de versant sont fréquemment voisines (0,48, thème "Etat du sol") et cette colocalisation, déjà mise en évidence par le traitement A dès 50 m de distance permet d'affiner l'interprétation en indiquant que l'essentiel de ces zones est en connexion directe. Les décharges minières non contrôlées et les mines sont intimement associées (0,83, thème "Nature du sol"). La végétation de maquis ligno-herbacé est contigüe soit à une végétation éparse (0,49), soit à un maquis dense paraforestier (0,86), montrant des systèmes végétaux qui peuvent être liés à la dégradation du milieu. Les colocalisations inter-thèmes indiquent aussi que les pistes sensibles sont très liées à la nature latéritique du sol (0,52), de même qu'à l'érosion de versant (0,56) ou aux mines (0,52). La savane et les végétations sur substrat volcano-sédimentaire semblent quant à elles présenter des affinités avec les alluvions (respectivement 0,64 et 0,5 ; 0,47 en colocalisation de taille 3) bien que ces dernières ne soient pas constituées du substrat typique. Le maquis ligno-herbacé ou le maquis dense paraforestier et la forêt sur substrat ultramafique sont nettement associés aux harzburgites (respectivement 0,61 et 0,48), ce qui s'explique par le fort taux d'endémisme de ces systèmes, avec une tendance à la dégradation sur sol latéritique, associé aux maquis (colocalisation de taille 3 à 0,31).

Les compléments apportés par ces résultats préliminaires à l'étude de l'érosion sont significatifs. Par exemple, il est possible de caractériser de manière plus pertinente des associations fréquentes, tout en y associant une quantification des relations entre objets difficiles à obtenir par ailleurs. Les traitements A et B sont complémentaires, notamment par la comparaison entre les résultats de A et de B pour une colocalisation donnée, qui permet de discerner si les objets concernés sont uniquement contigus ou s'ils se trouvent dans le même périmètre sans contact direct, les implications en termes d'érosion étant différentes.

4. Les motifs spatio-temporels : S2P

Nous nous intéressons dans cette section à un type particulier de motifs, les motifs spatio-temporels S2P qui permettent d'étudier des évènements se déplaçant dans le temps et l'espace. Nous présentons l'indice de participation spatio-temporel qui nous permet d'identifier les motifs les plus pertinents et finalement, nous décrivons l'application de cette méthode pour le suivi des épidémies de dengue.

4.1. Les données

Dans cette approche, nous étudions plus particulièrement l'évolution dans le temps de chaque zone en fonction des zones voisines. Contrairement, aux données utilisées pour les colocalisations où les zones sont très hétérogènes et peuvent apparaître/disparaître, les zones considérées dans cette approche sont fixes et ne se superposent pas. Elles correspondent typiquement aux quartiers d'une ville ou à des régions.

Cette méthode sera illustrée sur des données portant sur le suivi des épidémies de dengue. Ces données ont été collectées dans une ville à Nouméa (Nouvelle Calédonie)[3] 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. Un exemple de motif recherché est celui de la propagation du virus dans les différents quartiers d'une ville. Le virus se propage à plusieurs endroits à la fois, tout en restant présent dans le quartier d'origine. Dans ce contexte, l'objectif est plutôt de rechercher comment se propage le virus de quartier en quartier, en fonction des caractéristiques environnementales statiques (p. ex. le nombre de piscines ou le nombre d'espaces verts) ou dynamiques (p. ex. la météo).

4.2. Le concept de motifs spatio-séquentiels : S2P

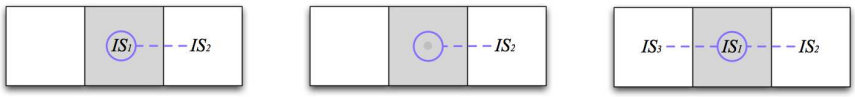
Item et Itemset. Soit un *item*, noté I, une valeur littérale associée à une dimension D_{A_i}, $I \in dom(D_{A_i})$. Un *itemset*, $IS = (I_1 I_2 \dots I_n)$ avec $n \leq p$, est un ensemble non vide d'*items* tel que $\forall i, j \in [1..n], \exists k, k' \in [1..p]$, $I_i \in dom(D_{A_k})$, $I_j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

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 DB est définie par : $In(IS, Z, T)$ 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 2). La notion d'*interaction* entre zones voisines est alors définie par :

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

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 $I_{ST} = IS_i \cdot IS_j$.

L'*opérateur de groupement* $n\text{-}aire$ noté *[ ]*, permet de regrouper une liste d'itemsets affectés par l'opérateur $\cdot$ (*à côté de*). Le symbole θ représente l'*absence d'itemsets* dans une zone. La figure 7 montre les 3 types d'itemsets spatiaux construits avec les notations précédentes. Les lignes pointillées représentent la dynamique spatiale. Dans la figure 7c, on observe dans la zone étudiée les ensembles d'évènements IS_1 et dans deux zones voisines différentes au même temps les ensembles d'évènements IS_3 et IS_2.



(a) $IS_1 \cdot IS_2$ (b) $\theta \cdot IS_2$ (c) $IS_1 \cdot [IS_2; IS_3]$

Figure 7. Représentation graphique des itemsets spatiaux

L'itemset spatial $I_{ST} = (T_m \cdot [V_l; P_m])$ signifie qu'il existe un temps t et une zone pour lesquelles l'événement T_m s'est produit et que l'évènement V_l et l'événement P_m se sont produits au même temps t 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 les événements T_m et P_l au même moment.

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 $(V_l) \subseteq (V_l 55)$.

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 8. Les flèches représentent la dynamique temporelle et les lignes pointillées représentent la relation de voisinage spatial entre itemsets.

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

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

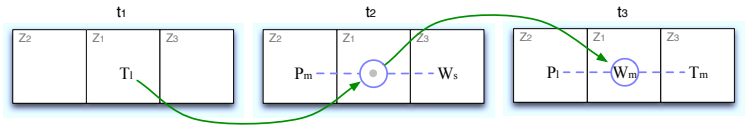


Figure 8. Exemple de dynamique spatio-temporelle

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 $(55) \subseteq (55 \cdot V_s)$.

Pour une zone spécifique Z, on note s_Z la séquence de données spatiales associée dans la base de données DB. 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 est ainsi défini comme le nombre de zones qui supportent s. 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)***. Nous présentons dans la section suivante une mesure de prévalence plus précise et adaptée.

4.3. 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, une nouvelle mesure de filtrage a été définie pour souligner la participation d'un élément dans une séquence spatiale. Elle est basée sur 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.

Ratio de participation spatial. Soit le motif spatio-séquentiel s et $I \in Dom(A)$ un item de s, le *ratio de participation spatial* de s par rapport à I noté $SPr(s, I)$ est défini par :

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

où $Supp(X)$ est le support classique qui est le nombre de zones contenant X.

Indice de participation spatial. Soit $s = \langle I_{ST_1}, I_{ST_2}, \ldots I_{ST_n} \rangle$ un motif spatio-séquentiel, l'*indice de participation spatial* de s noté $SPi(s)$ est le minimum de ses *ratios de participation spatiaux* :

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

Ratio de participation temporel. Soit le motif spatio-séquentiel 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.

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

Indice de participation temporel. Soit $s = \langle I_{ST_1}, I_{ST_2}, \ldots I_{ST_n} \rangle$ un motif spatio-séquentiel, l'*indice de participation temporel* de s noté $TPi(s)$ est le minimum de ses *ratios de participation temporels* :

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

L'*indice de participation spatio-temporel* d'un motif spatio-séquentiel s est défini comme le produit pondéré des deux mesures précédentes, noté $STPi(s)$:

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

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 *min_stpi*. Notons que le prédicat "STPI 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 est inclus dans s' ($s \preceq s'$), 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.

Pour extraire ces motifs, (Alatrista-Salas *et al.*, 2012) ont proposé un algorithme appelé *DFS-S2PMiner* qui suit une stratégie de recherche en profondeur *(depth-first search)*.

4.4. Résultats expérimentaux

DFS-S2PMiner a été développé en Java et testé sur le jeu de données réelles des données épidémiologiques. Notre approche est comparée avec la méthode la plus proche sémantiquement : l'algorithme DFS_Mine proposé par Tsoukatos (Tsoukatos, 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 sur Linux.

4.4.1. Résultats quantitatifs et performances

Une évaluation quantitative de notre approche a été réalisée. La figure 10 indique les temps d'exécution des deux algorithmes DFS_Mine (support classique) et

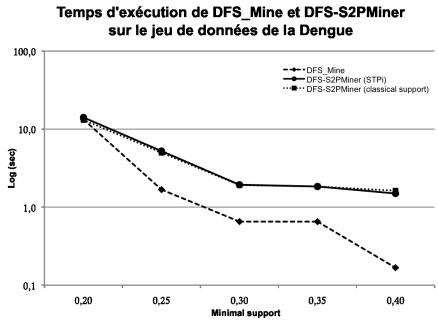


Figure 9. Temps d'exécution de DFS_Mine et DFS-S2PMiner

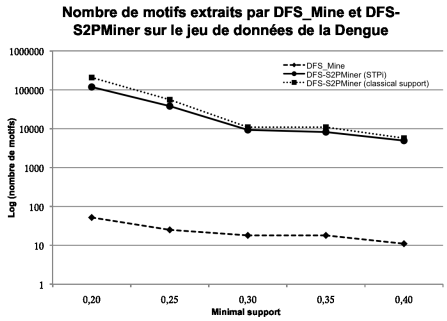


Figure 10. Nombre de motifs S2P extraits par DFS_Mine et DFS-S2PMiner

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 même si notre approche réalise 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.

4.4.2. Résultats qualitatifs et retours des experts

Une évaluation qualitative des résultats a été faite. Les motifs obtenus par notre approche ont été comparés 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 comme mesure d'intérêt, il n'est plus possible de 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. DFS_Mine extrait des motifs se produisant dans de nombreuses zones et à plusieurs dates. L'intérêt de notre proposition est de considérer le poids temporel des motifs extraits.

Les motifs spatio-sequentiels S2P 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 temporelle de la dengue dans le temps en fonction des caractéristiques décrivant les quartiers et leur voisinage. Les perspectives associées à ce travail sont essentiellement d'étendre la notion de voisinage à 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.

5. Conclusions générales et challenges

Dans cet article, nous avons présenté deux types de bases de données spatio-temporelles : celles permettant de travailler sur les trajectoires d'objets mobiles et celles permettant de manipuler les évènements localisés. Nous avons ensuite présenté une revue des méthodes de découvertes de motifs dans des bases de données spatio-temporelles (section 2.3). Nous avons ensuite présenté deux exemples de méthodes que nous avons adoptées et/ou développées et enfin appliquées sur des problématiques réelles (section 3 et 4). La première approche est une alternative à des méthodes d'analyse spatiale dans les données SIG et est basée sur la notion de motifs spatiaux (p. ex. les colocalisations). Ce nouveau domaine de motifs prend en compte la dimension spatiale en plus de la dimension thématique (d'attributs) et permet de décrire des corrélations ou des tendances en termes de caractéristiques mais aussi en termes de relations spatiales entre les objets étudiés. Les colocalisations (section 3) ont démontré leur intérêt et pertinence comme modèles descriptifs dans des problématiques traitant de l'analyse spatiale dans des données spatiales et SIG (où la dimension spatiale est importante) comme le montre le nombre de travaux étudiant ce type de motifs (cf. section 2.3). Ils permettent d'étudier les associations des caractéristiques d'objets spatialement (avec différentes relations spatiales) de manière globale. L'intégration de la dimension temporelle aux colocalisations a permis la définition d'un domaine de motifs spatio-temporels. Nous avons choisi de généraliser et d'étendre la méthode d'extraction des colocalisations dans une base de données SIG afin d'intégrer des contraintes thématiques permettant d'améliorer la pertinence des motifs et les performances de leur extraction. Mais ces motifs ne peuvent fournir que des informations du type "ensemble de caractéristiques souvent proches spatialement" (par leurs instances) le restent dans une période de temps. La deuxième approche, quant à elle, concerne l'analyse spatio-temporelle et est basée sur l'extension spatiale des techniques de fouille de données temporelles. Les motifs spatio-temporels $S2P$ introduits en section 4 en sont un exemple. Ils correspondent à une adaptation des motifs séquentiels classiques pour prendre en compte les relations spatiales. Contrairement

aux colocalisations les $S2P$ décrivent localement les évolutions temporelles d'événements (ou caractéristiques) en fonction de l'environnement voisin. L'application de ces motifs à deux jeux de données réelles (l'analyse spatiale de l'érosion et la propagation du virus de la dengue) a montré l'intérêt de ces motifs et les résultats fournis et leur interprétation par les experts semblent prometteurs.

Les challenges associés aux bases de données spatiales et spatio-temporelles sont multiples. Tout d'abord, la sémantique des motifs extraits doit être considérée pour présenter aux experts des motifs réellement adaptés à leurs besoins applicatifs. Dans cet article, nous avons présenté les motifs spatio-séquentiels qui sont un type de motifs basés sur la notion de séquence. Des motifs à structure plus complexe, comme les graphes attribués, peuvent s'avérer réellement performants dans les bases de données spatiales comme le montrent les travaux prometteurs de Pasquier *et al.* (2013) et Sanhes *et al.* (2013). Par ailleurs, les méthodes de fouille de données spatio-temporelles génèrent souvent de très nombreux motifs, parfois plus nombreux que les données initiales. Il est donc important de définir des mesures d'intérêts qui permettent aux experts de sélectionner les motifs les plus pertinents. Comme cela a été proposé dans la méthode basée sur les colocalisations, il est également nécessaire d'inclure dans le processus d'extraction, des connaissances du domaines (p. ex. méta-données, descriptions sémantiques, ontologies...) pour améliorer le passage à l'échelle mais également la qualité des motifs obtenus et leur interprétation. De plus, la définition de visualisations performantes adaptées à ces nouveaux motifs facilitera leur interprétation. Concernant les domaines d'application, beaucoup restent à explorer comme par exemple la fouille d'images où beaucoup de données existent mais peu de méthodes efficaces et passant à l'échelle ont été développées. Finalement, il existe une réel besoin de collaboration entre experts de la fouille de données et experts du domaine applicatif, collaboration qui reste la clef du succès du processus d'extraction de connaissances.

Remerciements

Les auteurs souhaitent remercier le projet ANR-COSINUS FOSTER (ANR-2012-COSI-012-01), la Direction des Affaires Sanitaires et Sociales de la Nouvelle Calédonie, de l'Institut Pasteur, de l'IRD et de l'UNC pour avoir fourni les données de la Dengue.

Bibliographie

Agrawal R., Imielinski T., Swami A. N. (1993). Mining association rules between sets of items in large databases. In P. Buneman, S. Jajodia (Eds.), *SIGMOD conference*, p. 207-216. ACM Press.

Agrawal R., Srikant R. (1995). Mining sequential patterns. In P. S. Yu, A. L. P. Chen (Eds.), *ICDE'95*, p. 3-14. IEEE Computer Society.

Alatrista-Salas H., Azé J., Bringay S., Cernesson F., Flouvat F., et al. (2011). Recherche de séquences spatio-temporelles peu contredites dans des données hydrologiques. *Revue des Nouvelles Technologies de l'Information (RNTI) - Numéro spécial "Qualité des Données*

et des Connaissances/Evaluation des Méthodes d'Extraction des Connaissances dans les Données", vol. E-22, p. 165-188.

Alatrista-Salas H., Bringay S., Flouvat F., Selmaoui-Folcher N., Teisseire M. (2012). The pattern next door: Towards spatio-sequential pattern discovery. In P.-N. Tan, S. Chawla, C. K. Ho, J. Bailey (Eds.), *PAKDD'12 (2)*, vol. 7302, p. 157-168. Springer.

Bogorny V., Valiati J. F., Silva Camargo S. da, Engel P. M., Kuijpers B., Alvares L. O. (2006). Mining maximal generalized frequent geographic patterns with knowledge constraints. In *ICDM'06*, p. 813-817. IEEE Computer Society.

Cao H., Mamoulis N., Cheung D. W. (2005). Mining frequent spatio-temporal sequential patterns. In *ICDM'05*, p. 82–89. Washington, DC, USA, IEEE Computer Society.

Cao H., Mamoulis N., Cheung D. W. (2007). Discovery of Periodic Patterns in Spatiotemporal Sequences. *IEEE TKDE*, vol. 19, n° 4, p. 453–467.

Celik M., Kang J. M., Shekhar S. (2007). Zonal co-location pattern discovery with dynamic parameters. In *ICDM'07*, p. 433-438. IEEE Computer Society.

Celik M., Shekhar S., Rogers J. P., Shine J. A. (2006). Sustained emerging spatio-temporal co-occurrence pattern mining: A summary of results. In *ICTAI'06*, p. 106-115. IEEE Computer Society.

Celik M., Shekhar S., Rogers J. P., Shine J. A. (2008). Mixed-drove spatiotemporal co-occurrence pattern mining. *IEEE TKDE*, vol. 20, n° 10, p. 1322-1335.

Desmier E., Flouvat F., Gay D., Selmaoui-Folcher N. (2011). A clustering-based visualization of colocation patterns. In *IDEAS'11*, p. 70-78. Springer.

DIMENC/SGNC, BRGM. (2005). *Carte géologique de la nouvelle-calédonie au 1/50 000.*

DTSI/SGT. (2008). *Cartographie de l'occupation du sol de la nouvelle-calédonie au 1/50 000.*

Fisher P., Laube P., Kreveld M., Imfeld S. (2005). Finding REMO - Detecting Relative Motion Patterns in Geospatial Lifelines. In *Developments in spatial data handling*, p. 201-215. Springer.

Flouvat F., Marchi F. D., Petit J.-M. (2009). The izi project: Easy prototyping of interesting pattern mining algorithms. In *PAKDD'09 workshops*, p. 1-15.

Flouvat F., Selmaoui-Folcher N., Gay D., Rouet I., Grison C. (2010). Constrained colocation mining: application to soil erosion characterization. In *SAC'10*, p. 1054-1059.

Giannotti F., Nanni M., Pinelli F., Pedreschi D. (2007). Trajectory pattern mining. In P. Berkhin, R. Caruana, X. Wu (Eds.), *SIGKDD'07*, p. 330-339. ACM.

Giannotti F., Pedreschi D. (Eds.). (2008). *Mobility, data mining and privacy - geographic knowledge discovery*. Springer.

Gudmundsson J., Kreveld M. J. van. (2006). Computing longest duration flocks in trajectory data. In R. A. de By, S. Nittel (Eds.), *GIS'06*, p. 35-42. ACM.

Gudmundsson J., Kreveld M. J. van, Speckmann B. (2004). Efficient detection of motion patterns in spatio-temporal data sets. In D. Pfoser, I. F. Cruz, M. Ronthaler (Eds.), *GIS'04*, p. 250-257. ACM.

Huang Y., Pei J., Xiong H. (2006). Mining co-location patterns with rare events from spatial data sets. *GeoInformatica*, vol. 10, n° 3, p. 239-260.

Huang Y., Shekhar S., Xiong H. (2004). Discovering colocation patterns from spatial data sets: A general approach. *IEEE TKDE*, vol. 16, nº 12, p. 1472-1485.

Huang Y., Zhang L., Zhang P. (2008, April). A framework for mining sequential patterns from spatio-temporal event data sets. *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, nº 4, p. 433-448.

Inokuchi A., Washio T., Motoda H. (2000). An apriori-based algorithm for mining frequent substructures from graph data. In D. A. Zighed, H. J. Komorowski, J. M. Zytkow (Eds.), *PKDD'00*, vol. 1910, p. 13-23. Springer.

Jensen C. S., Schneider M., Seeger B., Tsotras V. J. (Eds.). (2001). *Advances in spatial and temporal databases, 7th international symposium, sstd 2001, redondo beach, ca, usa, july 12-15, 2001, proceedings* (vol. 2121). Springer.

Jeung H., Yiu M. L., Zhou X., Jensen C. S., Shen H. T. (2008). Discovery of convoys in trajectory databases. *PVLDB*, vol. 1, nº 1, p. 1068-1080.

Kalnis P., Mamoulis N., Bakiras S. (2005). On discovering moving clusters in spatio-temporal data. In *SSTD'05*, vol. 3633, p. 364-381. Springer.

Li Z., Ding B., Han J., Kays R. (2010). Swarm: Mining relaxed temporal moving object clusters. *PVLDB*, vol. 3, nº 1, p. 723-734.

Lin J., Li Y. (2009). Finding structural similarity in time series data using bag-of-patterns representation. In M. Winslett (Ed.), *SSDBM'09*, vol. 5566, p. 461-477. Springer.

Mabit L., Selmaoui-Folcher N., Flouvat F. (2011). Modélisation de la dynamique de phénomènes spatio-temporels par des séquences de motifs. In A. Khenchaf, P. Poncelet (Eds.), *EGC'11*, vol. RNTI-E-20, p. 455-466. Hermann-Éditions.

Mamoulis N., Cao H., Kollios G., Hadjieleftheriou M., Tao Y., Cheung D. W. (2004). Mining, indexing, and querying historical spatiotemporal data. In *KDD'04*, p. 236–245. New York, NY, USA, ACM.

Mannila H., Toivonen H. (1997). Levelwise search and borders of theories in knowledge discovery. *Data Min. Knowl. Discov.*, vol. 1, nº 3, p. 241-258.

Mannila H., Toivonen H., Verkamo A. I. (1997). Discovery of frequent episodes in event sequences. *Data Min. Knowl. Discov.*, vol. 1, nº 3, p. 259-289.

Masseglia F., Cathala F., Poncelet P. (1998). The psp approach for mining sequential patterns. In J. M. Zytkow, M. Quafafou (Eds.), *PKDD'98*, vol. 1510, p. 176-184. Springer.

Nanni M., Pedreschi D. (2006, November). Time-focused clustering of trajectories of moving objects. *J. Intell. Inf. Syst.*, vol. 27, nº 3, p. 267–289.

Pasquier C., Sanhes J., Flouvat F., Selmaoui-Folcher N. (2013). Frequent pattern mining in attributed trees. In *PaKDD'13*, p. 26-37. Springer.

Pei J., Han J., Mortazavi-Asl B., Wang J., Pinto H., et al. (2004). Mining sequential patterns by pattern-growth: The prefixspan approach. *IEEE TKDE*, vol. 16, nº 11, p. 1424-1440.

Piatetsky-Shapiro G. (1991). Discovery, analysis, and presentation of strong rules. In *KDD'91*, p. 229-248. AAAI Press.

Qian F., He Q., He J. (2009). Mining spread patterns of spatio-temporal co-occurrences over zones. In O. Gervasi, D. Taniar, B. Murgante, A. Laganà, Y. Mun, M. L. Gavrilova (Eds.), *ICCSA'09 (2)*, vol. 5593, p. 677-692. Springer.

Rouet I., Gay D., Allenbach M., Selmaoui N., Ausseil A.-G., et al. (2009). Tools for soil erosion mapping and hazard assessment: application to new caledonia, sw pacific. In *MODSIM'09*, p. 1986–1992.

Sanhes J., Flouvat F., Pasquier C., Selmaoui-Folcher N., Boulicaut J.-F. (2013). Weighted path as a condensed pattern in a single attributed dag. In *IJCAI*.

Selmaoui-Folcher N., Flouvat F. (2011). How to use "classical" tree mining algorithms to find complex spatio-temporal patterns? In A. Hameurlain, S. W. Liddle, K.-D. Schewe, X. Zhou (Eds.), *DEXA'11 (2)*, vol. 6861, p. 107-117. Springer.

Shekhar S., Huang Y. (2001). Discovering spatial co-location patterns: A summary of results. In C. S. Jensen, M. Schneider, B. Seeger, V. J. Tsotras (Eds.), *SSTD'01*, vol. 2121, p. 236-256. Springer.

Srikant R., Agrawal R. (1996). Mining quantitative association rules in large relational tables. In H. V. Jagadish, I. S. Mumick (Eds.), *SIGMOD'96*, p. 1-12. ACM Press.

Termier A., Rousset M.-C., Sebag M. (2002). Treefinder: a first step towards xml data mining. In *ICDM'02*, p. 450-457. IEEE Computer Society.

Tsoukatos I., Gunopulos D. (2001). Efficient mining of spatiotemporal patterns. In C. S. Jensen, M. Schneider, B. Seeger, V. J. Tsotras (Eds.), *SSTD'01*, vol. 2121, p. 425-442. Springer.

Verhein F., Al-Naymat G. (2007). Fast mining of complex spatial co-location patterns using glimit. In *ICDM'07 workshops*, p. 679-684. IEEE Computer Society.

Wang J., Hsu W., Lee M.-L. (2005). Mining generalized spatio-temporal patterns. In L. Zhou, B. C. Ooi, X. Meng (Eds.), *DASFAA'05*, vol. 3453, p. 649-661. Springer.

Wang J., Hsu W., Lee M.-L., Wang J. T.-L. (2004). Flowminer: Finding flow patterns in spatio-temporal databases. In *ICTAI'04*, p. 14-21. IEEE Computer Society.

Wang L., Zhou L., Lu J., Yip J. (2009). An order-clique-based approach for mining maximal co-locations. *Inf. Sci.*, vol. 179, n° 19, p. 3370-3382.

Wang Y., Lim E.-P., Hwang S.-Y. (2006). Efficient mining of group patterns from user movement data. *DKE*, vol. 57, n° 3, p. 240-282.

Yang H., Parthasarathy S., Mehta S. (2005). A generalized framework for mining spatio-temporal patterns in scientific data. In R. Grossman, R. J. Bayardo, K. P. Bennett (Eds.), *KDD'05*, p. 716-721. ACM.

Yoo J. S., Bow M. (2009). Finding n-most prevalent colocated event sets. In T. B. Pedersen, M. K. Mohania, A. M. Tjoa (Eds.), *DaWaK'09*, vol. 5691, p. 415-427. Springer.

Yoo J. S., Shekhar S. (2006). A joinless approach for mining spatial colocation patterns. *IEEE TKDE*, vol. 18, n° 10, p. 1323-1337.

Zaki M. (2001, Jan/Feb). SPADE: an efficient algorithm for mining frequent sequences. *MLJ, Special issue on Unsupervised Learning*, vol. 42, n° 1/2, p. 31-60.

Zaki M. J. (2002). Efficiently mining frequent trees in a forest. In *KDD'02*, p. 71-80. ACM.

Zhang X., Mamoulis N., Cheung D. W., Shou Y. (2004). Fast mining of spatial collocations. In W. Kim, R. Kohavi, J. Gehrke, W. DuMouchel (Eds.), *KDD'04*, p. 384-393. ACM.