S2PViewer : un prototype de visualisation de motifs spatio-temporels

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

▶ **To cite this version:**

Hugo Alatrista Salas, Sandra Bringay, Frédéric Flouvat, Nazha Selmaoui-Folcher, Maguelonne Teisseire. S2PViewer : un prototype de visualisation de motifs spatio-temporels. Conférence Internationale de Géomatique et d'analyse spatiale (SAGEO), Sep 2013, Brest, France. hal-02599582

HAL Id: hal-02599582
https://hal.inrae.fr/hal-02599582

Submitted on 29 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

S2PViewer : un prototype de visualisation de motifs spatio-temporels

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

1. TETIS - Irstea
500, rue J. F. Breton, 34093 Montpellier, Cedex 5, FRANCE
prenom.nom@teledetection.fr

2. Lirmm
161, rue Ada, 34392 Montpellier, Cedex 5, FRANCE
prenom.nom@lirmm.fr

3. PPME – Université de la Nouvelle Calédonie
98851 Nouméa, Cedex, NOUVELLE CALEDONIE
prenom.nom@univ-nc.nc

RESUME.

Le volume des données collectées et stockées dans des bases de données spatio-temporelles augmente. Il devient donc crucial de fournir des synthèses permettant aux experts de mieux appréhender ces données afin de prendre des décisions pour agir immédiatement. Dans ce contexte, la fouille de données spatiales, qui permet d'identifier de nouveaux motifs dans ces bases, couplée à des méthodes de visualisation, facilite le travail des experts. La visualisation apporte notamment une forte valeur ajoutée pour la compréhension de dynamiques spatio-temporelles de motifs extraits. Dans cet article, nous présentons une nouvelle approche de visualisation dédiée aux motifs spatio-temporels. Notre approche a été validée sur une base de données contenant des informations sur le suivi épidémiologique de la dengue en Nouvelle Calédonie.

ABSTRACT.

The volume of stored data in spatiotemporal databases increases rapiditly. It becomes crucial to provide summaries from these data allow experts to make decisions and to act immediately on the results. In this context, spatial data mining, which helps to identify new patterns in these databases coupled with visualization methods, facilitates the work of experts. Visualization derived from data mining can provide great add value, often crucial for the understanding of some spatiotemporal dynamics. In this paper, we present a new visualization approach of spatiotemporal patterns. Our proposition has been tested on a database containing information about epidemiological surveillance of dengue in New Caledonia.

MOTS-CLES : Base de données spatio-temporelle, épidémies, fouille de motifs séquentiels.

KEYWORDS: Spatiotemporal databases, epidemics, sequential patterns mining.

1. Introduction

De nombreux phénomènes évoluent dans l'espace et le temps. La modélisation de ces phénomènes est souvent complexe, non seulement en raison de leur nature spatiale et temporelle, mais aussi du fait des interactions possibles entre les événements participant aux phénomènes. L'étude de la propagation d'une maladie vectorielle (e.g. la dengue) dans une ville en est un exemple typique. Les experts en santé publique savent que l'évolution de l'épidémie dépend de facteurs environnementaux (e.g. climat, proximité de zones infectées en points d'eau, mangroves...) et d'interactions entre les humains et le vecteur de transmission (e.g. le moustique qui transporte la maladie). Toutefois, l'impact de ces facteurs environnementaux et de ces interactions reste encore mal connu. Dans ce contexte, les méthodes d'Extraction de Connaissances à partir des Données (ECD) apportent des solutions via l'identification sans hypothèse *a priori* de relations entre variables et événements, caractérisées dans l'espace et dans le temps. Malheureusement, l'exploitation des connaissances extraites par l'expert est souvent limitée, car il lui est difficile de s'approprier ces nouvelles connaissances qui sont parfois tout aussi complexes à interpréter que les données initiales, notamment, lorsque l'on cherche à représenter des dynamiques spatiales et temporelles (Cao *et al*., 2011). La mise en place de méthodes et d'outils permettant de mieux restituer ces connaissances est donc un enjeu majeur.

Dans cet article, nous nous focalisons sur la fouille de données spatio-temporelles, et plus particulièrement, sur la restitution des connaissances obtenues aux experts. Pour cela, nous nous appuyons sur un nouveau type de motifs appelés *motifs spatio-séquentiels*, définis dans Alatrista-Salas *et al*. (2012). Ces motifs sont basés sur une extension des motifs séquentiels (Agrawal et Srikant, 1995), de façon à considérer ensemble les dimensions spatiale et temporelle. Nous proposons une approche de visualisation permettant aux experts de mieux appréhender les interactions spatiales et temporelles entre les différents facteurs représentés par ces motifs. À la différence des approches classiques, la méthode de visualisation retenue souligne les dynamiques spatio-temporelles, tout en prenant en compte l'environnement proche. Notre méthode de visualisation peut être appliquée à d'autres types de motifs, où les dimensions spatiales et/ou temporelles sont présentes, telles que les co-localisations (Shekhar et Huang, 2001) et les séquences temporelles de Tsoukatos et Gunopoulos (2001).

Nous ne proposons pas uniquement une méthode de visualisation de motifs spatio-temporels, mais tout un environnement permettant de faire une analyse détaillée de ces motifs à différentes échelles (des motifs globaux aux objets spatiaux locaux). Plus précisément, notre environnement de visualisation offre les avantages suivants :

- un affichage synthétique et schématique des motifs sous forme de graphes colorés (avec possibilité d'associer des icones aux nœuds). Trois visualisations sont proposées en fonction des besoins des experts et du type de motifs ;
- un affichage détaillé des zones et des dates où sont apparus les événements (i.e. des occurrences des motifs). À partir d'un motif, il est possible d'identifier les zones impactées sur une carte (et inversement). Une frise chronologique permet de visualiser les dates des évènements représentés par les motifs (avec deux niveaux de détails) ;
- des statistiques détaillées sur les zones (e.g. nombre d'habitants) et les caractéristiques temporelles des motifs (e.g. durée moyenne).

À notre connaissance, il n'existe pas d'approches de visualisation qui propose ce type de fonctionnalités associées à des motifs spatio-temporels. Notre proposition a été appliquée au suivi épidémiologique de la dengue dans la ville de Nouméa. Les premiers retours des experts professionnels de santé confirment l'intérêt d'une telle plateforme.

L'organisation de cet article est la suivante : un état de l'art complet est présenté dans la Section 2. Ensuite, dans la Section 3, nous introduisons les motifs spatio-séquentiels ainsi que la problématique soulevée. La Section 4 détaille notre approche de visualisation. Nous présentons le prototype de visualisation et son application aux données de la dengue dans la Section 5, pour conclure et présenter les perspectives en Section 6.

2. Etat de l'art

La visualisation d'informations est un thème très étudié ces dernières années dans diverses disciplines informatiques, et en particulier pour les résultats obtenus après l'étape de fouille de données.

Dans un contexte général, les techniques de visualisation ont été largement discutées dans la littérature (voir par exemple, Tufte, 1983 ; Peuquet, 1994 ; Bertin, 2010). Ces auteurs insistent, entre autres, sur l'importance de l'affichage visuel des informations de façon à rendre plus facile l'interprétation des résultats obtenus. Un graphisme visuellement attrayant, montrant l'information à interpréter, est plus intéressant et souvent plus efficace en terme d'interprétation qu'un affichage immédiat de quelques chiffres ou une représentation purement textuelle (Tufte, 1983). Ces auteurs donnent également des indications sur les spécifications techniques à prendre en compte lors de la représentation visuelle des données. Ils arrivent à la conclusion que la visualisation graphique d'informations doit être précise, claire et efficace. Ces trois caractéristiques doivent être accompagnées de conventions d'ordre technique comprenant : la sélection des couleurs, le choix des formes, les polices, la forme et le remplissage des lignes, le rangement des espaces de design et bien d'autres. Par ailleurs, Keim (2002) propose une classification des techniques de visualisation, selon le type de données à visualiser, e.g. données unidimensionnelles, bidimensionnelles, hiérarchiques, etc.

Plus spécifiquement, les systèmes de visualisation de motifs sont souvent associés à de nouvelles techniques d'extraction de connaissances appliquées à différents domaines. En effet, il est très difficile d'aborder le problème de la visualisation des connaissances extraites par la fouille de données sans aborder la méthode qui précède le système de visualisation. Plusieurs approches ont été proposées pour visualiser des motifs séquentiels. Par exemple, Pak Chung Wong *et al*. (2000) ont appliqué une technique d'extraction de motifs séquentiels aux données textuelles et l'ont accompagnée d'un prototype de visualisation permettant l'analyse des motifs obtenus sur des grands corpus. Subasic et Berendt (2008) ont proposé une méthode et un outil de visualisation pour cartographier et interagir avec les publications scientifiques postées sur le Web en utilisant des méthodes de fouille de textes. Plus récemment, Sallaberry *et al*. (2011) ont présenté un cadre pour la modélisation et la visualisation de motifs séquentiels permettant d'identifier les associations et les relations hiérarchiques entre des données associées à des gènes humains.

Dans le cas des données spatio-temporelles, un état de l'art très complet a été fait par Adrienko *et al*. (2003). Ils présentent un inventaire des techniques d'exploration visuelles existantes en fonction du type de données et des méthodes de fouille de données utilisées. Trois types de données spatio-temporelles ont été étudiés : (1) les données représentant des changements existentiels des objets spatiaux, e.g. occurrence, (2) les données reflétant des changements dans les propriétés spatiales des objets, et ; (3) les données représentant des variations temporelles des différents attributs thématiques.

Par ailleurs, Bertini et Lalanne (2010) étudient le rôle de la visualisation et des techniques de fouille de données sur le processus d'extraction de connaissances à partir des données (ECD). Ils distinguent trois catégories de techniques : (1) les techniques fondamentalement visuelles mais qui exigent l'exécution d'un processus de calcul avant la visualisation des résultats ; (2) les techniques où la fouille de données est l'étape prédominante et les résultats sont montrés en s'appuyant sur un système de visualisation, et ; (3) les techniques où la fouille de données et la visualisation sont totalement intégrées (il est impossible de distinguer lequel des deux processus joue un rôle prédominant). Ils ont aussi proposé des extensions possibles à ces travaux.

Du coté applicatif, Ping *et al*. (2008) ont visualisé des motifs représentant des changements au niveau d'objets géographiques (e.g. expansion de villes). Les motifs extraits étaient des règles d'association spatio-temporelles. Ils les ont représenté en prenant en compte quatre caractéristiques : date, durée, ordre et fréquence. Plus récemment, Burauskaite-Harju *et al*. (2012) ont proposé un ensemble de méthodes permettant, entre autres, la visualisation des dépendances spatiales entre des épisodes des précipitations temporellement synchronisés.

Les travaux étudiant le problème de la visualisation dans le processus ECD sont donc nombreux. Toutefois, contrairement à notre approche, ils ne sont pas adaptés à des motifs spatio-temporels complexes faisant à la fois intervenir les dynamiques spatiale et temporelle et un environnement proche. Un motif tel que *« la présence de cas de dengue dans une région est souvent précédée de températures élevées dans une zone située près de réservoirs d'eau »* pourrait difficilement être observé via ces méthodes.

3. Motifs spatio-séquentiels : concepts et définitions

Dans cette section, nous présentons la méthode de fouille de données permettant d'extraire des *motifs spatio-séquentiels* (cf. Alatrista-Salas *et al*., 2012).

3.1. Définitions préliminaires

Une base de données spatio-temporelle $stDB$ est un ensemble d'informations structurées contenant des composantes géographiques (e.g. quartiers, rivières, etc.), des composantes temporelles (e.g. des dates) et des données décrivant les composantes géographiques à un temps donné. Plus formellement, une base de données spatio-temporelle est définie comme un triplet $stDB = (D_T, D_S, D_A)$ où D_T est la dimension temporelle, D_S est la dimension spatiale and $D_A = \{D_{A_1}, D_{A_2}, \dots, D_{A_p}\}$ est un ensemble des dimensions d'analyse associées aux attributs. Le Tableau 1 montre un exemple de base de données spatio-temporelle et sera utilisée dans le reste de cette section.

Tableau 1. Exemple d'une base de données spatio-temporelle (où b=bas, m=moyen, h=haut)

Zone	*Date*	*Température*	*Précipitations*	*Vent*	*Nombre de cas de dengue*
Z_1	12/03/2013	T_b	P_m	V_m	--
Z_1	13/03/2013	T_m	P_m	V_b	3
Z_1	14/03/2013	T_b	P_m	V_m	11
Z_2	12/03/2013	T_m	P_m	V_m	4
Z_2	13/03/2013	T_b	P_m	V_b	2
Z_2	14/03/2013	T_b	P_b	V_m	--

La dimension temporelle est associée à un domaine de valeurs noté $dom(D_T) = \{T_1, T_2, \dots, T_t\}$ où $\forall i \in [1..t]$, T_i est souvent appelé estampille temporelle et $T_1 < T_2 < \dots < T_t$.

Chaque dimension D_{A_i} ($\forall i \in [1..p]$) appartenant à la dimension d'analyse est associée à un domaine de valeurs noté par $dom(A_i)$. Dans ce domaine, les valeurs peuvent être ordonnées ou non.

La dimension spatiale est associée à un domaine de valeurs noté $dom(\{D_S) = \{Z_1, Z_2, \dots, Z_l\}$ où $\forall i \in [1..l], Z_i$ est une zone.

Définition 1. Item et Itemset

Soit un événement I, traditionnellement appelé *item*, une valeur littérale pour la 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 $\forall\, i, j \in [1..n]$, $\exists k, k' \in [1..p]$, $I_i \in dom(D_{A_k})$, $I_j \in dom(D_{A_{k'}})$ et $k \neq k'$.

Définition 2. Séquence d'itemsets

Une séquence d'itemsets S est une liste ordonnée, non vide, d'itemsets notée $< s_1\ s_2\ \dots\ s_q >$ où s_j est un itemset.

Tous les événements produits dans une même zone sont regroupés et triés par date. Par exemple, considérons les événements produits dans la zone *Z1* du *12/03/2013* au *14/03/2013* selon la séquence $S = <(T_b\ P_m\ V_m)(T_m\ P_m\ V_b 3)(T_b\ P_m\ V_m\ 11)>$ indiquée dans le Tableau 1. Ceci signifie que hormis les événements T_b, P_m, et V_m qui se sont produits au même temps dans la zone *Z1*, i.e. lors de la même transaction, les autres événements de la séquence se sont produits dans deux autres dates dans la même zone. Une n-séquence est une séquence composée de n items, dans notre exemple, S est une 10-séquence.

Une séquence $< s_1\ s_2\ \dots\ s_q >$ est une sous-séquence d'une autre séquence $< s'_1\ s'_2\ \dots\ s'_m >$ s'il existe des entiers $< i_1\ i_2\ \dots\ i_j\ \dots\ i_q >$ tels que $1 \leq i_1 \leq i_2 \leq \dots \leq i_p \leq \dots \leq i_m$ Par exemple, la séquence $<(AB)(C)>$ est inclue dans la séquence $<(AB)(CD)>$ car $(AB) \subseteq (AB)$ et $(C) \subseteq (CD)$.

3.2. Motifs spatio-séquentiels

Nous allons étendre les définitions présentées dans la Section 3.1 de façon à prendre en compte la dynamique spatiale des données. Pour cela, nous définissons la relation *dans* entre une zone Z et un itemsets IS comme l'occurrence de l'itemset IS dans la zone Z au temps t dans la base de données $stDB$. Plus formellement :

$$\begin{cases} dans(IS, Z, t) = vrai\ si\ IS\ apparait\ dans\ stDB\ pour\ la\ zone\ Z\ au\ temps\ t \\ dans(IS, Z, t) = faux\ sinon \end{cases}$$

Ensuite, nous définissons la notion de voisinage entre zones qui peut être de différente nature, par exemple, les zones qui se trouvent à une certaine distance ou celles qui partagent une frontière, etc. Dans ce travail, deux zones Z_i et Z_j sont voisines si :

$$\begin{cases} voisin(Z_i, Z_j) = vrai\ si\ Z_i\ et\ Z_j\ partagent\ une\ frontière \\ voisin(Z_i, Z_j) = faux\ sinon \end{cases}$$

Définition 3. Itemset spatial

Soient IS_i and IS_j deux itemsets, IS_i et IS_j sont spatialement proches si $\exists\ Z_f, Z_g \in dom(D_S)$ et $\exists\ t \in dom(D_T)$ tel que $dans(IS_i, Z_f, t) \wedge dans(IS_j, Z_g, t) \wedge voisin(Z_f, Z_g)$ est vrai. Deux itemsets IS_i et IS_j qui sont spatialement proches, forment un itemset spatial noté $I_{ST} = IS_i \cdot IS_j$. Les itemsets spatiaux représentent un profil (non-exhaustif par rapport à l'ensemble des attributs disponibles) d'une zone et de son environnement proche.

Pour alléger les notations, nous introduisons un opérateur de groupement d'itemsets associés à l'opérateur · (voisin) et noté []. Le symbole θ représente l'absence d'itemsets dans une zone. Par exemple, l'itemset spatial $(\theta \cdot [B; C])$ peut être interprété comme deux événements différents apparus dans deux zones proches à une zone où aucun événement ne s'est produit à un moment donné.

Définition 4. Association entre zone, itemset spatial et le temps

Soit $I_{ST} = IS_i \cdot IS_i$ un itemset spatial, $Z \in dom(D_S)$ une zone et $t \in dom(D_T)$ une estampille temporelle, nous définissons la relation *vérifier* qui représente la présence de l'itemset spatial I_{ST} dans Z au temps t comme suit :

$$\begin{cases} vérifier(I_{ST}, Z, t) = vrai\ si\ dans(IS_i, Z, t) = vrai\ et\ Z' \in dom(D_S)\ tel\ que \\ \qquad\qquad\qquad\qquad voisin(Z, Z') = vrai\ et\ dans(IS_j, Z', t) = vrai \\ vérifier(I_{ST}, Z, t) = false\ sinon \end{cases}$$

Définition 5. Inclusion d'itemsets spatiaux

Un itemset spatial $I_{ST} = IS_i \cdot IS_i$ est inclus, noté par $\subseteq$, dans autre itemset spatial $I'_{ST} = IS'_k \cdot IS'_l$, si et seulement si $IS_i \subseteq IS'_k$ et $IS_j \subseteq IS'_l$.

Puis, nous modélisons la notion d'évolution d'événements dans les zones en prenant en compte leur relation de voisinage via la notion de séquence spatiale.

Définition 6. Séquence spatiale

Une séquence spatiale ou simplement *2S* est une liste ordonnée d'itemsets spatiaux notée par $s = < I_{ST_1} I_{ST_2} \ldots I_{ST_m} >$ où $I_{ST_i}, I_{ST_{i+1}}$ satisfaisant la contrainte de séquentialité temporelle, i.e. $i \in [1\,..\,m-1]$.

Ensuite, nous définissons une mesure d'élagage appelée *support absolu* pour des séquences spatiales définie comme le nombre de zones contenant la séquence étudiée et satisfaisant les contraintes de proximité des *itemsets spatiaux*. Plus formellement, cette mesure peut être définie comme suit :

Définition 7. Support absolu d'une séquence spatiale

Soit la *2S* $s = < I_{ST_1} I_{ST_2} \ldots I_{ST_m} >$, le support absolu de s représenté par $supp_{abs}(s,\ stDB)$ est défini comme le nombre de zones de la base de données *stBD* qui vérifient s, autrement dit :

$$supp_{abs}(s, stBD) = \left| \begin{matrix} Z \in dom(D_S)\ tel\ que\ \forall i \in [1..m], \exists t_i \in dom(D_T) \\ et\ vérifier(I_{ST_i}, Z, t_i) = vrai \end{matrix} \right|$$

De la même manière, nous définissons le support relatif dénoté par $supp_{rel}(s,\ stDB)$ pour une *2S* s comme le ratio entre le support absolu et le nombre de zones total.

Définition 8. Support relatif d'une séquence spatiale

Soient la *2S* $s = < I_{ST_1} I_{ST_2} \ldots I_{ST_m} >$, $supp_{abs}(s,\ stDB)$ le support absolu de s et $|dom(D_S)|$ le nombre total de zones de la base de données spatio-temporelle *stDB*, le support relatif de s est défini par :

$$supp_{rel}(s, stBD) = \left| \frac{supp_{abs}(s, stBD)}{|dom(D_S)|} \right|$$

Définition 9. Motif spatio-séquentiel

Soient la *2S* s et σ le support minimum spécifié par l'utilisateur, si $supp_{rel}(s, stDB) \geq \sigma$ alors s est une *2S* fréquente appelée motif spatio-séquentiel (S2P) où $supp_{rel}$ est le support relatif d'une séquence spatiale. Si la valeur de σ est proche de *1*, des motifs dit « rares » seront extraits. Contrairement, si σ est proche de *0*, des motifs « triviaux » seront restitués.

4. S2Pviewer : un environnement pour visualiser des motifs spatio-temporels

Dans cette section, nous présentons une nouvelle approche de visualisation des motifs spatio-séquentiels précédemment définis. Cette méthode a été intégrée dans un prototype de visualisation appelé S2PViewer qui permet de visualiser la dynamique spatiale (voisinage proche) ainsi que la dynamique temporelle. S2PViewer est basé sur Javascript, JQuery et D3 et le prototype est accessible sur le site http://datamining.univ-nc.nc/.

4.1. Une visualisation globale des motifs sous forme de graphes colorés

Comme indiqué dans la section précédente, un motif spatio-séquentiel (S2P) est une séquence d'itemsets spatiaux. Notre approche de visualisation doit permettre, entre autres, d'avoir un aperçu général de ces motifs, de donner un sens à ce qu'ils représentent et de trouver un motif pertinent parmi les nombreux motifs inintéressants (Bertini et Lalanne, 2010). Pour représenter visuellement de tels motifs, il faut prendre en compte deux dynamiques : la dynamique spatiale, représentée par les itemsets spatiaux (l'opérateur spatial) et la dynamique temporelle, représentée par l'aspect séquentiel.

Un itemset spatial représente l'état courant d'une zone (ses évènements ou caractéristiques) ainsi que celui de ses voisins proches, à un instant donné. Trois types d'itemsets spatiaux peuvent être représentés à l'aide des opérateurs tels que · (voisin), *[]* (groupement) et le symbole θ (absence). La Figure 1 illustre ces trois cas en utilisant une représentation à base de graphe.

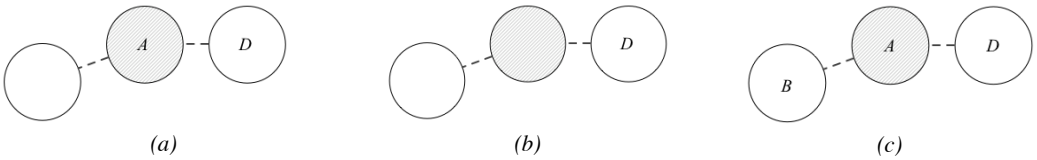

Figure 1. Représentation graphique des itemsets spatiaux (a) A·D (b) θ ·D (c) A· [B ; D]

Dans la Figure 1, chaque cercle représente une zone, et les étiquettes représentent les évènements associés à la zone (i.e., les itemsets). Dans les trois figures, la zone centrale colorée est la zone étudiée. Les lignes pointillées représentent le voisinage spatial. Cette représentation du voisinage spatial a été définie par Peuquet D.J., (1994) comme une représentation de la « contiguïté spatiale ». La longueur et l'angle des lignes pointillées n'ont pas de signification particulière.

Les motifs spatio-séquentiels décrivent l'évolution temporelle d'une zone et de son environnement proche à différentes estampilles temporelles. Cette évolution temporelle est représentée par une succession d'itemsets spatiaux. Le S2P illustré dans la Figure 2 est composé de trois estampilles temporelles. La première est composée de deux événements *A* et *B* apparus dans la zone étudiée. Après, il n'existe aucun événement dans la zone étudiée mais *B* et *C* sont apparus dans des zones voisines. Enfin, *P* est apparu dans la zone étudiée et *Q* et *R* sont apparus autour. Les flèches pleines représentent la dynamique temporelle.

Comme on peut le constater, cette représentation respecte le postulat énoncé par Tufte (1983) : la quantité d'informations (dimensions) affichées ne dépasse pas le nombre de dimensions présente dans les données. En effet, la temporalité est représentée par un arc orienté, qui symbolise une succession d'événements, et l'aspect spatial (voisinage) est représenté par des arcs en pointillés.

Afin d'enrichir cette représentation, nous utilisons la taille et la couleur des nœuds pour donner visuellement des informations supplémentaires à l'expert. La taille du nœud est directement proportionnelle au nombre d'items (i.e., d'évènements) qu'il représente. La couleur est utilisée pour

identifier visuellement certains évènements d'intérêt particulier pour l'expert (e.g., la présence de dengue). Une solution efficace pour représenter les types d'entités (e.g., présence de dengue ou absence de dengue) est l'utilisation des régions fermées colorées (Ware, 2004). Nous avons choisi des cercles de deux couleurs différentes : le vert et le rouge car ce sont des couleurs opposées dans le modèle de couleurs appelé *color opponent-process model*. En effet, un œil sain les distingue parfaitement, contrairement à d'autres paires de couleurs (Ware, 2004). Aussi, la couleur rouge est souvent employée, par convention, pour représenter les maladies, le danger, etc., contrairement à la couleur verte qui est souvent utilisée pour représenter des personnes saines, des zones non polluées, etc.

Un point reste à préciser par rapport à cette représentation : comment positionner les éléments du graphe dans l'espace ? En effet, un même graphe peut être affiché de différentes façons en fonction de la position des nœuds et des arcs. La Figure 3 présente trois exemples de configuration. Une technique classique pour afficher des graphes est d'utiliser les algorithmes basés sur les forces (Kaufmann et Wagner, 2001). Cette approche positionne les nœuds du graphe de façon à limiter les croisements entre arêtes. Un des avantages de cette approche est de permettre à l'utilisateur de déplacer les nœuds du graphe. Lorsqu'un nœud est déplacé, le graphe est alors réorganisé automatiquement à la volée (c.f. Figure 3a).

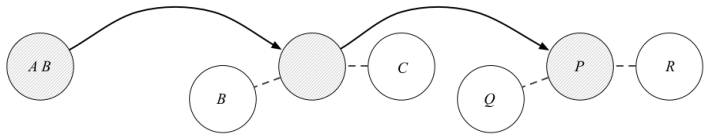

Figure 2. Représentation graphique du S2P <(AB) (θ · [B ; C]) (P · [Q ; R])>

Notre objectif premier est d'avoir un affichage qui soit le plus intuitif possible pour les utilisateurs, i.e., un affichage facilitant la compréhension du motif. Pour un motif spatio-séquentiel représentant l'évolution dans le temps d'une zone (et de son environnement), un affichage sous forme de frise chronologique (ou ligne du temps) semble être un affichage naturel. Cet affichage revient simplement à positionner les nœuds (de gauche à droite) sur un arc orienté de rayon fixe. Les nœuds représentant l'environnement proche sont répartis « autour » de cette ligne de temps reliés par des lignes pointillées. Cette représentation est illustrée par la Figure 3b. Une représentation plus classique du temps sous la forme d'une ligne droite n'a été pas retenue car elle ne permet pas de positionner autant d'éléments dans une même fenêtre.

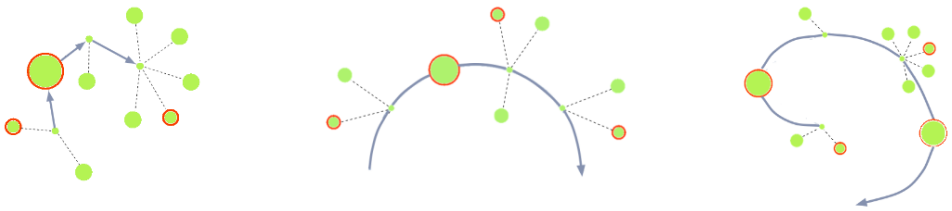

Figure 3. Trois types de visualisation (a) graphe, (b) arc, (c) spirale

Par ailleurs, lorsque le motif à afficher est très long et contient de nombreux itemsets spatiaux, cet affichage sous forme d'arc peut ne pas suffire. Dans ce contexte, nous avons privilégié une visualisation basée sur des coordonnées polaires – au lieu des coordonnées cartésiennes. En effet, le ratio de l'arc peut être incrémental et dépend de la taille du motif (Bertin, 2010). L'angle théta représente la succession ou le temps. Le ratio de l'arc étant incrémental, cette approche permet de visualiser de longs motifs sous la forme d'une spirale d'Archimède (cf. Figure 3c).

Les conventions citées précédemment sont intégrées dans notre prototype. Nous laissons à l'utilisateur le choix du type de positionnement (graphe « dynamique », arc et spirale). En effet, chaque approche a ses avantages et inconvénients. La représentation en arc est la plus naturelle, mais elle n'est pas adaptée si les motifs sont longs. La représentation en spirale est le plus adaptée à longs motifs, mais elle est moins intuitive. La représentation en graphe « dynamique » (basée sur les algorithmes de forces) est probablement la moins adaptée à notre problème. Par contre, elle pourra être particulièrement intéressante pour afficher des motifs à n dimensions, par exemple, pour représenter la propagation d'une épidémie.

La représentation d'un motif donne une vision globale de celui-ci (disposition des événements au cours du temps et localisation dans l'espace). Elle peut être complétée par une visualisation à une granularité plus fine (celle des zones et des temps où est apparu le motif), comme nous le décrivons dans la sous-section suivante.

4.2. Vers une visualisation d'information à différents niveaux

Notre approche de visualisation des motifs S2P se veut une approche globale avec des possibilités de visualisation d'informations détaillées localement sur les éléments d'un motif (e.g. où et quand apparaît ce motif ?).

La visualisation de la dynamique temporelle a pour objectif de valoriser l'information concernant la durée de l'occurrence d'un S2P et où il est apparu. Deux points ont été pris en compte au moment de la représentation de la dynamique temporelle : (1) pour chaque motif, une synthèse contenant les périodes, dates du début et de fin de l'apparition du motif dans les zones concernées sera représentée en utilisant des blocs de couleurs différentes. Actuellement, la visualisation des zones impactées pour le motif sélectionné est présentée à l'aide d'une liste ; (2) des statistiques concernant la durée maximale, minimale et moyenne d'un motif, par rapport aux zones où le motif est impacté, seront présentées également. Ces valeurs calculées « à la volé » permettront à l'expert d'avoir une référence des caractéristiques concernant la durée d'apparition des motifs sur les zones associées.

Figure 4. Exemple de représentation de la dynamique temporelle d'une séquence spatiale

Pour présenter la dynamique temporelle (voir Figure 4), les conventions à prendre en compte pour sa conception sont différentes de celles considérées pour la représentation de la dynamique spatiale. En effet, si l'on utilise des coordonnées polaires, alors une déformation des motifs les plus éloignés sera perçue (ils seront plus longs). De plus, comme cette visualisation n'est pas centrale mais indicative, une représentation en utilisant des coordonnées cartésiennes est privilégiée. Donc, nous allons représenter l'apparition d'un motif avec un bloc représentant les dates du début et fin d'un motif dans une zone.

Par exemple, soit le S2P $s =< (\boldsymbol{A})(\boldsymbol{B}) >$ et la séquence caractérisant une zone $S =< (\boldsymbol{AC})(D)(D)(\boldsymbol{BC}) >$, nous représentons l'apparition de la séquence spatiale s dans S avec un bloc qui commence au temps t_0 et finis au temps t_3. Si un S2P apparaît de multiples fois dans une même zone, il sera tracé plusieurs fois avec la même couleur. De même, si le motif apparaît dans plus d'une zone, de multiples blocs seront tracés en utilisant des différentes couleurs.

4.3. Le processus d'analyse spatio-temporelle avec S2PViewer

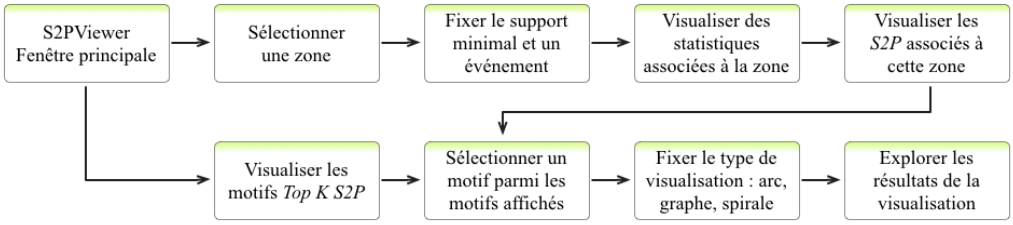

Figure 5. Diagramme de flux du processus d'analyse de S2P

Notre prototype de visualisation est basé sur la technique décrite par Andrienko *et al*., (2003) appelée « interaction de cartes » dans laquelle, des événements spatio-temporels sont « mappés » en utilisant une carte. La Figure 5 montre le diagramme de flux du processus d'analyse spatio-temporelle des S2P. Le prototype de visualisation est composé de trois étapes qui se traduisent par trois vues interactives dans notre interface :

1. La première vue – fenêtre principale – permet de sélectionner une zone. La Figure 6a montre l'exemple des données épidémiologiques (données dengue en Nouvelle Calédonie) avec comme zones spatiales les quartiers de Nouméa. Dans cette vue, une seule zone pourra être

sélectionnée à chaque fois que l'on désire afficher des motifs la décrivant. Une fois la zone sélectionnée, un support minimal doit être choisi (le support représente la probabilité d'apparition de ces motifs dans la base de données). Enfin dans cette vue, l'utilisateur devra sélectionner un événement considéré comme plus intéressant pour lui, sinon un événement (par exemple *dengue*) est sélectionné par défaut (voir Figure 6d).

2. La deuxième vue montre des informations statistiques décrivant la zone choisie dans la vue précédente (voir Figure 6b). Notamment, nous avons ajouté de l'information concernant la population de la zone (recensement de l'année 2004), le nombre de cas de dengue par année, et des informations sur les zones voisines, i.e. ceux qui partagent une frontière commune avec la zone sélectionnée auparavant.

3. En fin (Figure 6c), la troisième vue montre les S2P appartenant à la zone sélectionnée ayant un support supérieur ou égal au support fixé dans la première vue. Cette vue interactive permet, la sélection d'un motif, la visualisation graphique de ce motif ainsi que la visualisation de la dynamique temporelle associée (voir Figure 7 et 8). Si aucune zone n'a été sélectionnée, les vingt plus longs motifs seront automatiquement visualisés.

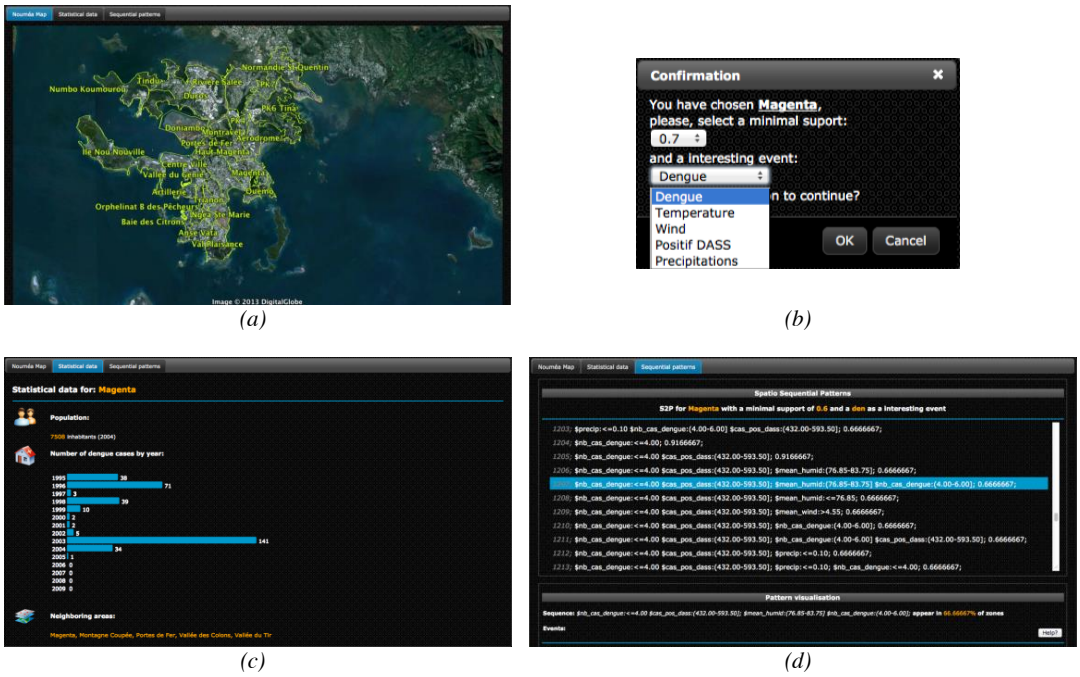

Figure 6. Les 4 vues du prototype de visualisation : (a) sélection d'un quartier, (b) sélection du support minimal et d'un événement intéressant, (c) visualisation des informations associées au quartier sélectionné, et (d) visualisation des 20 plus longs motifs ou des S2P associés au quartier sélectionné

5. Cas d'étude : analyse d'une épidémie de dengue

5.1. Contexte

Notre approche est générique et peut être appliquée à différents problèmes comme l'érosion des sols, de pollution des rivières, etc. Dans cet article, nous avons testé notre prototype sur une base de données spatio-temporelle contenant des données épidémiologiques de suivi de la dengue. Ces données ont été collectées dans la ville de Nouméa (Nouvelle Calédonie) sur un territoire divisée en *32* quartiers couvrant *45,7 km²*. Cette division spatiale a été proposée par la *Direction des Affaires Sanitaires et Sociales en Nouvelle Calédonie* (DASS). Cet ensemble de données contient des informations associées à des données démographiques, entomologiques, météorologiques, ainsi que des données de planification urbaine et des données médicales.

Afin d'obtenir des données catégorielles (données séparables dans des catégories qui s'excluent mutuellement), une discrétisation a été faite pour transformer les données continues en données nominales à l'aide de la méthode de discrétisation des *fréquences égales*. Les données sont stockées

selon trois classes de valeurs dans lesquelles ils sont classées : basse, moyenne et haute. Ces classes ont été discutées avec les experts. Le Tableau 2 montre des exemples de motifs extraits par notre méthode. Par exemple, le dernier motif peut être interprété par : dans *60%* des zones, les événements *haute température* et *haute humidité* apparaissent dans une zone voisine avant la présence de *peu de poubelles* et de l'apparition d'un certain nombre de *cas de dengue* (inferieur à 6) dans la zone d'étude.

Tableau 2. Exemple de S2P

Motif spatio-séquentiel	*Support*
< (mean_temper:>23.55 ihre_index:>34.82)(nb_cas_dengue:<=6.00 mean_temper:<=23.55)(mean_humid:(76.85-83.30]) >	0,7
< (waste_container:<=39.00 nb_cas_dengue:<=6.00)(θ·[comunity_gather:<=20.00 ; nb_cas_dengue:<=6.00 ; graveyard:<=2.00]) >	0,6
< (θ·[mean_temper:>23.55 ; mean_humid:>83.30])(waste_container:<=39.00 nb_cas_dengue:<=6.00) >	0,6

Sur les données de la dengue, nous avons souligné dans Alatrista-Salas *et al*., (2012), que notre approche permet de trouver des motifs tels que : *peu de dépôts d'eau, peu de précipitations et faible présence de locaux publics suivi par quelques cas de dengue, peu de précipitations et du vent*. Cependant, notre approche trouve également des motifs plus complexes tels *que peu de dépôts d'eau, peu de précipitations et faible présence de locaux publics suivi par peu de dépôts d'eau et beaucoup de précipitations dans deux zones voisines, suivi par présence de dengue dans une zone voisine à la zone d'étude*. L'exemple souligne de la richesse des motifs extraits par notre approche en mettant en évidence l'influence des zones voisines.

Notre outil de visualisation permet de visualiser tout type de séquences d'itemsets (ensemble d'événements) décrivant une zone (objet géo-référencé) et son environnement proche (des zones voisines) quelque soit l'application, e.g. érosion des sols, pollution des rivières, épidémies, etc.

5.2. Analyse sémantique des motifs obtenus

La Figure 7 représente visuellement un S2P. Comme a été décrit dans la Section 4.1, chaque cercle symbolise un ensemble d'événements apparaissant à un moment donné. La taille des cercles reflète le nombre d'évènements caractérisant la zone. L'absence d'événement dans la zone choisie (i.e. θ) est représentée par un cercle de petite taille. L'arc de cercle représente le temps et les lignes pointillées représentent la dimension spatiale (à côté de).

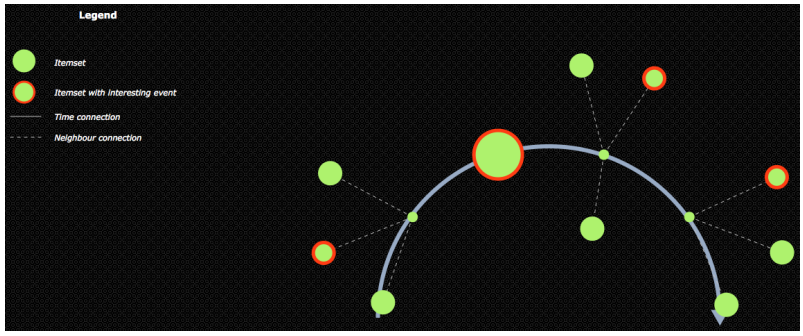

Figure 7. Représentation d'un S2P à l'aide du S2PViewer

La couleur rouge sert à identifier les itemsets qui contiennent l'événement *dengue* ou tout autre événement choisi lors de la première étape (voir Figure 6d). Cette caractéristique permet d'identifier facilement, d'un côté, la position (i.e. l'estampille temporelle) dans le motif S2P où se trouve l'événement intéressant, et d'un autre coté, les autres événements appartenant au même itemset.

Comment l'ont mentionné Andrienko *et al*., (2003), il est nécessaire de différencier les événements temporaires des événements durables. Les changements de caractéristiques décrivant une zone n'ont pas la même interprétation s'ils apparaissent dans des périodes courtes ou longues. Dans ce contexte, la représentation de la dynamique temporelle de notre approche permet d'identifier aussi bien la durée d'un motif que la périodicité de son apparition dans toutes les zones. Par exemple, la Figure 8 illustre un S2P qui apparaît périodiquement dans les quartiers *PK7, PK6 Tina, Ducos et Rivière Salée*. Cette information peut devenir cruciale pour les experts quand une épidémie est sur le point de se déclencher.

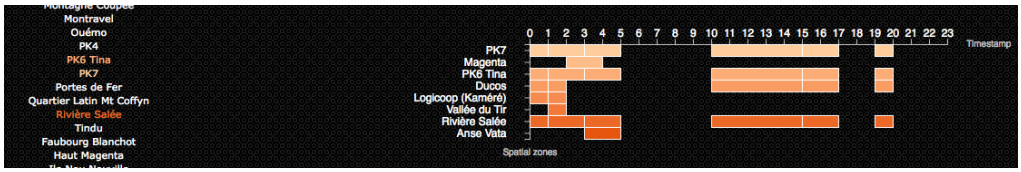

Figure 8. Représentation de la dynamique temporelle

5.3. Temps de réponse et validation pour les experts

Notre prototype de visualisation prend comme fichier d'entrée les motifs obtenus lors de l'étape de fouille de données et transforme cette information en objets à visualiser. Cette transformation ne consomme pas beaucoup de ressources. Au cours de nos expérimentations, nous avons affiché facilement plus de 3000 motifs en quelques secondes. Ce temps démontre la réactivité de notre prototype. Cette expérimentation met également en avant la nécessité de développer des fonctionnalités pour filtrer et/ou classer les motifs. En effet, il est difficilement envisageable de faire interpréter autant de motifs à un expert. Ce problème a motivé l'intégration dans notre prototype d'une fonctionnalité permettant de visualiser les « top-k » motifs (c.à.d. les *k* meilleurs motifs).

Concernant la validation de notre prototype, nous avons travaillé en partenariat avec des experts en santé public de l'Institut Louis Pasteur, de l'IRD et de la Direction des Affaires Sanitaires et Sociales de la Nouvelle Calédonie (DAAS) afin de recueillir des expertises, d'analyser les besoins et ainsi de résoudre les problèmes de lisibilité et d'utilisabilité de notre prototype. Lors de ces collaborations, nous nous sommes intéressés aux données à fouiller et aux besoins des utilisateurs finaux de S2PViewer pour réunir les informations concernant : (1) l'utilité, i.e. le prototype doit permettre d'aboutir à un résultat et ce résultat doit être pertinent pour ses objectifs, et ; (2) l'utilisabilité, i.e. le prototype doit permettre de réaliser une action rapidement et efficacement. Nous avons ainsi conçu trois types de visualisation, soit en arc, en graphe ou en spirale, lesquelles ont été développées suite aux remarques des experts.

6. Conclusion et perspectives

Dans cet article, nous avons proposé une nouvelle approche de visualisation permettant la restitution de motifs spatio-séquentiels (S2P). Ces motifs décrivent l'évolution d'un ensemble de caractéristiques "spatialisées" dans le temps en prenant en compte l'environnement voisin. Cet outil de visualisation permet de souligner la dynamique spatiale et temporelle des motifs extraits. Nous avons testé notre méthode d'extraction et de visualisation des S2P sur des jeux de données réelles dont celles associées au suivi épidémiologique de la dengue en Nouvelle Calédonie. Les résultats montrent l'intérêt de l'approche pour extraire et visualiser efficacement des S2P très riches.

Les perspectives associées à ce travail sont nombreuses. À terme, nous utiliserons des icones avec différentes couleurs pour représenter et identifier des événements. En effet, la représentation des événements par des icones avec des couleurs variant selon l'intensité (e.g. haute → rouge, basse → bleu) semble tout à fait pertinente. Cette idée a été soutenue par les experts en épidémiologie impliqués dans le projet. Nous voudrions aussi réduire le nombre de motifs à visualiser. Pour cela, nous sommes à la recherche de méthodes permettant, par exemple, de montrer les plus longues séquences ou celles contentant un événement spécifique à une position donnée (e.g. l'apparition de la dengue dans le dernier itemset). Une autre perspective consiste à étendre au spatial des mesures spatialisées comme la moindre contradiction temporelle (cf. Alatrista-Salas *et al*., 2011). Par exemple, pour trouver des motifs qui se contredisent ou non selon les zones ou des périodes temporelles ou des événements. Par ailleurs, des améliorations sur la visualisation de motifs sont en cours de développement, notamment la représentation des motifs en utilisant une vue plus abstraite permettant à l'utilisateur de s'orienter dans les données. Nous voudrions également proposer le prototype en application mobile pour « smartphones ». En effet, il sera possible de récupérer la position géo-référencée de l'utilisateur, identifier la zone et afficher les informations et les S2P appartenant à cette zone automatiquement.

Remerciements :

Nous tenons à remercier la Direction des Affaires Sanitaires et Sociales de la Nouvelle Calédonie, l'Institut Pasteur, l'IRD et l'Université de Nouvelle Calédonie (Convention 2010) pour les données de suivi

épidémiologique de la dengue, ainsi qu'Arnaud Salaberry et Pierre Accorsi pour leur contribution dans la conception et le développement de l'outil de visualisation.

Bibliographie

Agrawal R., et Srikant R., (1995). *Mining sequential patterns*. Data Engineering, International Conference on, 0:3, 1995.

Andrienko N., Andrienko G., Gatalsky P., (2003). *Exploratory spatio-temporal visualization: an analytical review*. Journal of Visual Languages & Computing, Volume 14, Issue 6, p. 503-541, 2003.

Alatrista-Salas H., Bringay S., Flouvat F. Selmaoui-Folcher N., et Teisseire M., (2012). *The Pattern Next Door: Towards Spatio-sequential Pattern Discovery*. Advances in Knowledge Discovery and Data Mining - 16th Pacific-Asia Conference PAKDD'12, p. 157-168, 2012.

Alatrista-Salas H., Cernesson F., Bringay S., Azé J., Flouvat F. Selmaoui-Folcher N., et Teisseire M., (2011). *Recherche de séquences spatio-temporelles peu contredites dans des données hydrologiques*. Revue des Nouvelles Technologies de l'Information, RNTI-E-22, p. 165-188, 2011

Bertini E. et Lalanne D., (2010). *Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery*. ACM SIGKDD Explor. Newsletter, p. 9-18, 2010.

Bertin, J., (2010). *Semiology of Graphics: Diagrams, Networks, Maps*. Economic & Social Research Institute, 2010.

Burauskaite-Harju A., Grimvall A., Walther A., Achberger C. et Chen D., (2012). *Characterizing and visualizing spatio-temporal patterns in hourly precipitation records*. Journal of Theoretical and Applied Climatology, Vol. 109, p. 333-343, 2012.

Cao L., Zhang H., Zhao Y., Luo D. et Zhang C., (2011). *Combined mining : Discovering informative knowledge in complex data*. Systems, Man, and Cybernetics, IEEE Transactions on, 41(3), p. 699-712, 2011.

Kaufmann M. et Wagner D., (2001). *Drawing Graphs: Methods and Models*. Lecture Notes in Computer Science, Springer, 2001.

Keim D.A., (2002). *Information Visualisation and Visual Data Mining*. Visualisation and computer graphics, IEEE Transaction on, 7(1), p. 100-107, 2002

Pak Chung Wong, Cowley, W., Foote, H., Jurrus, E., Thomas, J., (2000). *Visualizing sequential patterns for text mining*. InfoVis 2000. IEEE Symposium on, p.105-111, 2000.

Peuquet D.J., (1994). *It's about Time: A Conceptual Framework for the Representation of Temporal Dynamics in Geographic Information Systems*. Annals of the Association of American Geographers , Vol. 84, No. 3, pp. 441-461, 1994.

Ping Y., Xinming T., et Shengxiao W., (2008). *Dynamic cartographic representation of spatio-temporal data*. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B2, p. 7-12, 2008

Sallaberry A., Pecheur N., Bringay S., Roche M., Teisseire M., (2011). *Sequential patterns mining and gene sequence visualization to discover novelty from microarray data*. Journal of Biomedical Informatics, Volume 44, Issue 5, p. 760-774, 2011.

Shekhar S. et Huang Y., (2001). *Discovering spatial co-location patterns a summary of results*. Advances in Spatial and Temporal Databases, p. 236-256, 2001.

Subasic, I. et Berendt, B., (2008). *Web Mining for Understanding Stories through Graph Visualisation*. ICDM '08. Eighth IEEE International Conference on, p. 570-579, 2008.

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

Tufte, E.R., (1983). *The visual display of quantitative information*. The Visual Display of Quantitative Information, Graphics Press, 1983.

Ware, C., (2004). *Information Visualization: Perception for Design*. Interactive Technologies, Elsevier Science, 2004.

View publication stats