Vers une extraction et une visualisation des co-localisations adaptées aux experts

Frédéric Flouvat*, Nazha Selmaoui-Folcher*,**, Dominique Gay*,**

*Pôle Pluridisciplinaire de la Matière et de l'Environnement (PPME)
** Equipe de Recherche en Informatique et Mathématiques (ERIM)
Université de la Nouvelle-Calédonie,
BP R4, F-98851 Nouméa, Nouvelle-Calédonie
{frederic.flouvat, nazha.selmaoui, dominique.gay}@univ-nc.nc

Résumé. Une des tâches classiques en fouille de données spatiales est l'extraction de co-localisations intéressantes dans des données géo-référencées. L'objectif est de trouver des sous-ensembles de caractéristiques booléennes apparaissant fréquemment dans des objets spatiaux voisins. Toutefois, les relations découvertes peuvent ne pas être pertinentes pour les experts, et leur interprétation sous forme textuelle peut être difficile. Nous proposons, dans ce contexte, une nouvelle approche pour intégrer la connaissance des experts dans la découverte des co-localisations, ainsi qu'une nouvelle représentation visuelle de ces motifs. Un prototype a été développé et intégré dans un SIG. Des expérimentations on été menées sur des données géologiques réelles, et les résultats validés par un expert du domaine.

1 Introduction

La fouille de données spatiales a pour objectif l'extraction de connaissances intéressantes, utiles, inattendues et cachées dans des données spatiales. Elle a de nombreuses applications en gestion de l'environnement, en sécurité publique, dans les transports, ou le tourisme. Un des principaux défis en fouille de données spatiales est de découvrir et délivrer aux experts du domaine des connaissances utiles et interprétables (cf Cao (2008)). Bien que beaucoup d'algorithmes et de méthodes aient été proposés, cela reste encore un problème ouvert. Dans ce contexte, la fouille de données guidée par le domaine (*domain driven data mining*) vise à fournir des solutions aux experts pour passer d'une découverte de connaissance centrée sur les données et les contraintes techniques, à une découverte de connaissance centrée sur l'expert.

Une des tâches classiques en fouille de données spatiales est l'extraction de co-localisations intéressantes dans des données géo-référencées. L'objectif est de trouver des sous-ensembles de caractéristiques booléennes apparaissant fréquemment dans des objets spatiaux voisins. Plusieurs approches ont été proposées pour extraire des co-localisations telles que Koperski et Han (1995); Shekhar et Huang (2001); Huang et al. (2004); Yoo et Shekhar (2006); Bogorny et al. (2006). Toutefois, les relations découvertes peuvent ne pas être pertinentes pour les experts car

déjà connues. De plus, les co-localisations sont présentées aux experts sous forme textuelle, ce qui rend difficile leur interprétation.

Face à ces problèmes, nous proposons une nouvelle approche pour intégrer la connaissance du domaine dans la découverte des co-localisations, ainsi qu'une nouvelle représentation visuelle de ces motifs. La prise en compte de la connaissance du domaine se fait grâce à l'intégration de contraintes thématiques et spatiales dans le processus d'extraction. Les résultats obtenus sont donc plus pertinents pour les experts, tout en améliorant les performances de l'extraction. Le système de visualisation proposé s'appuie sur une représentation cartographique des co-localisations intégrée dans un SIG. Cette représentation simple, concise et intuitive des co-localisations prend également en considération la nature spatiale des objets sous-jacents et les pratiques des experts. Un prototype a été développé et intégré dans un SIG. Des expérimentations on été menées sur des données géologiques réelles, et les résultats validés par un expert du domaine.

La section 2 présente un rapide état de l'art sur l'extraction de motifs spatiaux et leur visualisation. La notion de co-localisation est présentée en détail et généralisée en section 3. La section 4 présente nos propositions afin d'avoir une découverte de co-localisations adaptée aux besoins des experts. Nous présentons ensuite une application de ces travaux au problème de l'érosion (section 5). Pour finir, nous concluons et donnons quelques perspectives à ce travail.

2 Etat de l'art

2.1 Extraction de motifs spatiaux

Deux approches ont été identifiées par Huang et al. (2004) pour l'extraction de motifs spatiaux : l'approche orientée transactions et l'approche orientée événements.

La première s'appuie sur une transformation des données spatiales en données transactionnelles, permettant ainsi l'utilisation d'algorithmes classiques d'extraction d'itemsets. Koperski et Han (1995) s'appuient sur cette approche pour extraire des règles d'association dans des bases de données géographiques en se focalisant sur une caractéristique prédéfinie. Leur méthode énumère les voisinages de la caractéristique spatiale étudiée afin de "matérialiser" un ensemble de transactions correspondant aux instances de celles-ci. Un algorithme d'extraction d'itemsets est ensuite appliqué sur ces transactions. Cette extraction permet ainsi de trouver les co-localisations liées à la caractéristique de référence. Bogorny et al. (2006) ont étendu ce travail en introduisant des contraintes dans une phase de prétraitements. Ces contraintes sont des associations déjà identifiées comme étant non-intéressantes pour les experts.

La deuxième approche se focalise sur les objets et leur relation de voisinage. Cette approche a été proposée par Shekhar et Huang (2001) et a notamment été étendue par Huang et al. (2004); Yoo et Shekhar (2006). L'objectif de leur approche est de trouver tous les sous-ensembles de caractéristiques spatiales souvent proches, appelés co-localisations. Contrairement à l'approche précédente, toutes les caractéristiques et relations de voisinage sont considérées, et les données ne sont pas modifiées. Une mesure d'intérêt a été introduite pour filtrer les co-localisations les plus importantes. Grâce à l'anti-monotonie de ce prédicat, un algorithme par niveau a ensuite été utilisé pour extraire les solutions.

De manière plus générale, un grand nombre de travaux ont étudié l'intégration de contraintes dans les algorithmes d'extraction d'itemsets. Pour les motifs spatiaux, à notre connaissance,

seul le travail de Bogorny et al. (2006) s'est intéressé à ce problème. Toutefois, leur travail s'appuie sur la première approche qui ne prend en compte que partiellement les relations spatiales, et nécessite un pré-traitement des données.

2.2 Visualisation des motifs

Plusieurs systèmes, tels que Brunk et al. (1997); Keim (2002), ont été proposés pour visualiser des résultats de fouille de données. A titre d'exemple, MineSet de Brunk et al. (1997) est un système interactif pour la fouille de données intégrant des modules de visualisation (statistique, arbre, graphique, carte). Chaque algorithme de fouille de données (e.g. modèle Bayésien, arbre de décision ou règles d'association) est couplé à un outil de visualisation. Les règles sont, par exemple, représentées dans un espace à deux dimensions avec pour axes les itemsets en partie gauche et droite des règles, et pour valeur une barre représentant la confiance de la règle.

Andrienko et Andrienko (1999) se sont intéressés à la visualisation des données spatiales et des résultats de la fouille de données. Pour les informations non-géographiques tels que les arbres ou les règles, le système construit des liens dynamiques entre la carte et les rapports (i.e. les résultats de l'extraction sous forme textuelle). Lorsque que le curseur de la souris est positionné sur un noeud de l'arbre ou sur une règle, ces liens mettent en valeur les objets correspondants sur la carte (et inversement).

Plus récemment, Leung et al. (2008) ont étudié la visualisation des itemsets fréquents. Ils ont développé un système appelé WiFIsViz permettant de visualiser les itemsets fréquents sous forme de graphes orthogonaux. Les itemsets sont placés dans un espace à deux dimensions, où l'axe des abscisses représente les articles et l'axe des ordonnées la fréquence. Un itemset X est représenté par une ligne horizontale connectant des noeuds, où chaque noeud représente un article de X. Les itemsets partageant des préfixes communs sont fusionnés, ce qui améliore la visualisation.

A notre connaissance, aucune des solutions existantes n'a été conçue pour visualiser des motifs spatiaux de manière simple, concise et intuitive pour les experts, tout en prenant en compte la nature spatiale des objets sous-jacents.

3 Extraction de co-localisations intéressantes

Dans cette section, nous allons tout d'abord présenter la notion de co-localisation définie dans Shekhar et Huang (2001); Huang et al. (2004), puis la généraliser et l'étendre grâce à un cadre théorique existant.

3.1 Présentation du concept de co-localisation

Soient $\mathcal{F} = \{f1, f2, ..., fk\}$ un ensemble de caractéristiques booléennes et $\mathcal{O} = \{o_1, o_2, ..., o_n\}$ un ensemble d'objets spatiaux. Une **co-localisation** est un sous-ensemble de caractéristiques de $\mathcal{F}$ associées à des objets spatiaux appartenant à $\mathcal{O}$. Ces co-localisations représentent des caractéristiques apparaissant fréquemment dans des objets voisins. Nous définissons la fonction Θ pour représenter formellement l'association entre les objets et les caractéristiques :

$$\forall o \in \mathcal{O}, \exists! f \in \mathcal{F}, \Theta(o) = f$$

Afin de simplifier les notations, l'objet o_i de $\mathcal{O}$ ayant la caractéristique f sera noté f_i. La figure 2 représente un exemple d'objets spatiaux et de caractéristiques associées. L'ensemble des caractéristiques $\mathcal{F}$ est $\{A, B, C, D, E\}$. L'ensemble des objets spatiaux $\mathcal{O}$ est $\{o_1, o_2, ..., o_{12}\}$. La relation $\Theta(o_9) = A$ est notée A_9.

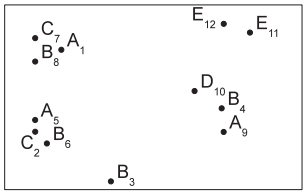

FIG. 1 – *Représentation graphique des objets spatiaux et de leur caractéristique*

Dans l'exemple de la figure 2, $ABCE$ est une co-localisation. Toutefois, toutes les co-localisations ne sont pas intéressantes. En effet, il n'existe aucun ensemble de quatre objets voisins ayant respectivement les caractéristiques A, B, C et E. Il convient donc d'introduire un certain nombre de notions afin de déterminer les co-localisations intéressantes.

Une **instance** I d'une co-localisation C, par rapport à une relation de voisinage $\mathcal{R}$ fixée, est un sous-ensemble d'objets de $\mathcal{O}$ ayant pour caractéristiques celles de C, et respectant deux à deux la relation spatiale $\mathcal{R}$. Elle vérifie donc les propriétés suivantes :

- $\forall o \in I, \Theta(o) = f \; avec \; f \in C$
- $|I| = |C|$
- $\forall o, q \in I, \mathcal{R}(o, q) = vraie$

Sur la figure 2, l'ensemble d'objets $\{A_1, B_8, C_7\}$ est une instance de la co-localisation $\{A, B, C\}$ (A_1, B_8, et C_7 voisins d'après $\mathcal{R}$). Nous dirons qu'un ensemble d'objets *vérifie* une co-localisation C par rapport à une relation de voisinage $\mathcal{R}$, lorsque ces objets constituent une instance de C. Par exemple, l'ensemble d'objets $\{A_1, B_8, C_7\}$ vérifie la co-localisation $\{A, B, C\}$, alors que $\{A_1, B_4, C_7\}$, $\{A_1, B_4, D_{10}\}$ ou $\{A_1, B_4\}$ ne la vérifient pas (figure 2).
De la même manière, un objet $o \in \mathcal{O}$ *participe* à une co-localisation C s'il appartient à une instance de C. Par exemple, les objets A_1, B_8, et C_7 participent chacun à la co-localisation $\{A, B, C\}$, alors que des objets tels que B_3 ou E_{12} n'y participent pas.

La **table d'instances** d'une co-localisation C, notée $TI_{\mathcal{R}}(\mathcal{O}, C)$ est l'ensemble des instances de C. Plus formellement, on a

$$TI_{\mathcal{R}}(\mathcal{O}, C) = \{I \subseteq \mathcal{O} \mid I \text{ est une instance de } C \text{ par rapport à } \mathcal{R}\}$$

Sur l'exemple de la figure 2, la table d'instances de $\{A, B, C\}$ est $TI_{\mathcal{R}}(\mathcal{O}, ABC) = \{ \{A_1, B_8, C_7\}, \{A_5, B_6, C_2\} \}$ et la table d'instances de $\{B, D\}$ est $TI_{\mathcal{R}}(\mathcal{O}, BD) = \{ \{B_4, D_{10}\} \}$.

Le **ratio de participation** d'une caractéristique f dans une co-localisation C, noté $pr_{\mathcal{R}}(\mathcal{O}, C, f)$, correspond à la fraction des objets de $\mathcal{O}$ ayant la caractéristique f et participant à la co-localisation C sur le nombre total d'objets ayant la caractéristique f.

$$pr_{\mathcal{R}}(\mathcal{O}, C, f) = \frac{|\{o \in I \mid I \in TI_{\mathcal{R}}(\mathcal{O}, C) \text{ et } \Theta(o) = f\}|}{|TI_{\mathcal{R}}(\mathcal{O}, f)|}$$

Sur l'exemple de la figure 2, $pr_{\mathcal{R}}(\mathcal{O}, \{A, B, C\}, A) = 2/3$, $pr_{\mathcal{R}}(\mathcal{O}, \{A, B, C\}, B) = 1/2$ et $pr_{\mathcal{R}}(\mathcal{O}, \{A, B, C\}, C) = 1$.

A partir de ces dernières définitions, Huang et al. (2004) ont introduit la notion d'**index de participation**, noté pi, pour évaluer la fréquence et la validité d'une co-localisation dans un jeu de données :

$$pi_{\mathcal{R}}(\mathcal{O}, C) = \min_{\forall f \in C}(pr_{\mathcal{R}}(\mathcal{O}, C, f))$$

Le **problème de l'extraction des co-localisations** peut donc être formulé de la manière suivante : *Soient $\mathcal{F}$ un ensemble de caractéristiques et $\mathcal{O}$ un ensemble d'objets spatiaux. Etant donnés une relation de voisinage $\mathcal{R}$ et un seuil $\sigma \in [0, 1]$, l'objectif est de trouver l'ensemble des co-localisations $C \subseteq \mathcal{F}$ telles que $pi_{\mathcal{R}}(\mathcal{O}, C) \geq \sigma$, avec pi la fonction utilisée pour calculer l'index de participation.*

3.2 Extension du cadre formel des co-localisations

Dans cette section, nous généralisons l'extraction des co-localisations grâce au cadre théorique de Mannila et Toivonen (1997), ce qui permettra par la suite d'intégrer des contraintes du domaine.

Le cadre théorique de l'extraction de motifs intéressants. Nous rappelons ici le cadre théorique défini dans Mannila et Toivonen (1997) pour les problèmes de découverte de motifs intéressants.

Soient une base de données d, un langage fini $\mathcal{L}$ représentant des motifs (au sens large) ou des sous-groupes des données, et un prédicat Q permettant d'évaluer si un motif $\varphi \in \mathcal{L}$ est "intéressant" dans d. Supposons en outre qu'une relation de spécialisation/généralisation, notée $\preceq$, est définie sur les éléments de $\mathcal{L}$. On dira que φ est plus général (resp. plus spécifique) que θ, si $\varphi \preceq \theta$ (resp. $\theta \preceq \varphi$). Nous supposons que le prédicat Q est anti-monotone (resp. monotone) par rapport à l'ordre $\preceq$. Ce qui signifie que pour chaque $\theta, \varphi \in \mathcal{L}$ tels que $\varphi \preceq \theta$, on a : $Q(d, \varphi)$ est faux (resp. vrai) $\Longrightarrow Q(d, \theta)$ est faux (resp. vrai).

La tâche de fouille de données consiste à extraire les motifs intéressants de d relativement à $\mathcal{L}$ et Q, appelé théorie de $d, \mathcal{L}, Q$, et définie par $Th(\mathcal{L}, d, Q) = \{\varphi \in \mathcal{L} \mid Q(d, \varphi)$ est vrai$\}$.

Généralisation et extension des co-localisations. L'extension des co-localisations au cadre générique précédent se fait de la manière suivante :
La base de données d correspond à une base de données géographiques composée d'un ensemble d'objets spatiaux $\mathcal{O}$ associés à une caractéristique de $\mathcal{F}$. Les éléments du langage $\mathcal{L}$ sont l'ensemble des co-localisations pouvant être formées à partir des caractéristiques de $\mathcal{F}$. La relation d'ordre partiel $\preceq$ entre les éléments du langage est l'inclusion ensembliste. Le prédicat Q est le prédicat anti-monotone qui retourne vrai (resp. faux) si la co-localisation a un index de participation supérieure (resp. inférieur) à un seuil minimum donné. L'ensemble des co-localisations à rechercher, i.e. la théorie, est donc $Th(\mathcal{L}, d, Q) = \{C \subseteq \mathcal{F} \mid pi_{\mathcal{R}}(\mathcal{O}, C) \geq \sigma\}$.

L'intégration de la notion de co-localisation dans le cadre formel précédent permet de généraliser ce problème à un problème de découverte de co-localisations intéressantes par rapport à un prédicat booléen Q et une relation spatiale booléenne $\mathcal{R}$ quelconques. La découverte de co-localisations n'est donc plus limitée à la seule mesure d'index de participation. En effet, il devient possible d'extraire des co-localisations respectant n'importe quel prédicat anti-monotone ou monotone (ou conjonction de prédicats), et ceci sans impact sur les algorithmes.

4 Vers une découverte de co-localisations adaptée aux besoins des experts

Dans les domaines manipulant des données géographiques, un des principaux outils utilisé pour stocker et visualiser l'information est le Système d'Information Géographique (SIG). L'intérêt de ces systèmes pour les experts est d'avoir une vision thématique et cartographique des données conforme à leurs habitudes de travail. Dans ce contexte, nous proposons dans cette section l'intégration de contraintes thématiques et spatiales dans la découverte des co-localisations, ainsi qu'une nouvelle représentation cartographique des co-localisations intégrée au SIG. Les résultats de l'extraction sont ainsi plus pertinents pour les experts, et leur interprétation facilitée.

4.1 Intégration de la connaissance du domaine

Dans cette section, nous allons exploiter des contraintes expertes pour intégrer la connaissance du domaine dans le processus de fouille de données. Dans notre contexte, les contraintes vont représenter des relations déjà connues par les experts ou non intéressantes, et qui devront être éliminées des résultats. En d'autres termes, ces contraintes peuvent être perçues comme des règles d'exclusion. Dans la suite, nous nous focaliserons sur des contraintes anti-monotones pouvant être utilisées pendant l'extraction. 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. Notre travail prend en compte ce découpage de l'information en permettant de définir des contraintes sur les caractéristiques, les thèmes et les objets. Nous considérons plus particulièrement deux types de contraintes :

- les contraintes sur les caractéristiques et les thèmes des co-localisations
- les contraintes spatiales sur les objets

Les contraintes sur les caractéristiques et les thèmes des co-localisations. Ce premier type de contraintes a pour objectif de permettre à l'expert de spécifier finement des co-localisations à ne pas étudier. Par exemple, l'expert n'est pas intéressé par les relations entre les caractéristiques *érosion de versant* et *harzburgites*.

Plus formellement, soient $X \in 2^{\mathcal{F}}$ un ensemble de caractéristiques à ne pas étudier conjointement, le prédicat $QC_X : 2^{\mathcal{F}} \rightarrow \{vrai, faux\}$ définissant les co-localisations C à exclure est défini de la manière suivante : $QC_X(C) : |C \cap X| = |X|$. Avant d'étendre cette première définition de QC_X aux thèmes, définissons plus formellement la notion thème.

Définition 1: *Soient $\mathcal{T} = \{t1, t2, ..., tw\}$ un ensemble de thèmes et $\mathcal{F} = \{f1, f2, ..., fk\}$ un ensemble de caractéristiques.*
La fonction $\Phi : \mathcal{F} \rightarrow \mathcal{T}$ associe à une caractéristique un thème telle que $\forall f \in \mathcal{F}$,
$\exists! t \in \mathcal{T}, \Phi(f) = t$.
La fonction $\Phi^{-1} : \mathcal{T} \rightarrow 2^{\mathcal{F}}$ associe à un thème un ensemble de caractéristiques telle que $\forall t \in \mathcal{T}, \Phi^{-1}(t) = \{f \in \mathcal{F} \mid \Phi(f) = t\}$.

A partir de cette définition, il est possible d'étendre le prédicat QC_X afin d'éliminer de l'analyse les co-localisations contenant un ensemble de caractéristiques et/ou abordant un en-

semble de thèmes. Il devient par exemple possible d'exclure les co-localisations étudiant la caractéristique *sol non nu* conjointement avec le thème *Végétation*.

Définition 2: *Soient $C \subseteq \mathcal{F}$ une co-localisation et $(F,T) \in 2^{\mathcal{F}} \times 2^{\mathcal{T}}$ un couple constitué d'un ensemble F de contraintes sur les caractéristiques et d'un ensemble T de contraintes sur les thèmes.*

$$QC_{F,T}(C) : |C \cap F| + |(\bigcup_{\forall f \in F} \Phi(f)) \cap T| = |F| + |T|$$

Soit $ConstrColoc = \{(F_1,T_1),...(F_u,T_u)\}$ l'ensemble des contraintes sur les co-localisations définies par les experts, avec $(F_i,T_i) \in 2^{\mathcal{F}} \times 2^{\mathcal{T}}$, $1 \leq i \leq u$. Soit $PPI_{\mathcal{R},\sigma}$ le prédicat s'appuyant sur le participation index pour sélectionner les motifs intéressants. Le prédicat $Q_{\mathcal{R}}$ utilisé dans l'algorithme d'extraction de co-localisations intéressantes devient donc :

$$Q_{\mathcal{R}} = PPI_{\mathcal{R},\sigma}(\mathcal{O},C) \wedge (\bigwedge_{\forall (F,T) \in ConstrColoc} \neg QC_{F,T}(C))$$

Ce nouveau prédicat reste anti-monotone car composé d'une conjonction de prédicats anti-monotones. Il peut être utilisé directement dans les algorithmes d'extraction de co-localisations intéressantes pour élaguer l'espace de recherche.

Les contraintes spatiales sur les objets. Ce deuxième type de contraintes a pour objectif de 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)".

Plus formellement, soient Z une zone géographique, $I \subseteq \mathcal{O}$ un ensemble d'objets, et $(F,T) \in 2^{\mathcal{F}} \times 2^{\mathcal{T}}$ un couple constitué d'un ensemble de caractéristiques et de thèmes. Le prédicat $QO_{Z,F,T} : 2^{\mathcal{O}} \rightarrow \{vrai, faux\}$ défini les instances de co-localisations à exclure de la manière suivante : $QO_{Z,F,T}(I) : |\{o \in I \mid \Theta(o) \in F \cup \Phi^{-1}(T) \text{ et } In(o,Z) = vrai\}| > 0$, avec $\Phi^{-1}(T)$ l'ensemble des caractéristiques des thèmes contenus dans T et $In(o,Z)$ une fonction testant si o est dans la zone Z. Il est possible de généraliser cette définition à toute relation spatiale booléenne entre un objet et une zone prédéfinie.

Définition 3: *Soient $I \subseteq \mathcal{O}$ un ensemble d'objets, Z une zone géographique, $\mathcal{R}'$ une relation spatiale booléenne, et $(F,T) \in 2^{\mathcal{F}} \times 2^{\mathcal{T}}$ un couple constitué d'un ensemble de caractéristiques et de thèmes.*

$$QO_{Z,\mathcal{R}',F,T}(I) : |\{o \in I \mid \Theta(o) \in F \cup \Phi^{-1}(T) \text{ et } \mathcal{R}'(o,Z) = vrai\}| > 0$$

avec $\Phi^{-1}(T)$ l'ensemble des caractéristiques des thèmes contenus dans T

Contrairement aux contraintes sur les co-localisations, les contraintes sur les objets n'apparaissent pas directement dans le prédicat $Q_{\mathcal{R}}$ utilisé pour l'extraction des co-localisations intéressantes. Elles influencent le calcul de l'index de participation en diminuant le nombre d'instances de co-localisations étudiées. La définition des tables d'instances se trouve donc modifiée. Soit $ConstrObj = \{(Z_0,\mathcal{R}_0,F_0,T_0),...,(Z_v,\mathcal{R}_v,F_v,T_v)\}$ l'ensemble des triplets Zone-Relation-Caractéristiques-Thèmes à exclure. La table d'instances $TI_{\mathcal{R}}(\mathcal{O},C)$ d'une co-localisation C dans un ensemble d'objets $\mathcal{O}$ est :

$$TI_{\mathcal{R}}(\mathcal{O}, C) = \{I \subseteq \mathcal{O} \mid I \text{ est une instance de } C \text{ par rapport à } \mathcal{R} \text{ et } QO_{Z,\mathcal{R}',F,T}(I) = faux, \forall (Z, \mathcal{R}', F, T) \in ConstrObj\}$$

4.2 Une visualisation spatiale des co-localisations intégrée dans un SIG

L'intégration des contraintes du domaine permet d'avoir des co-localisations plus intéressantes pour l'expert, mais leur interprétation sous forme textuelle reste difficile. Dans ce contexte, notre objectif est de proposer une visualisation cartographique des co-localisations intégrée au SIG, respectant ainsi les pratiques des experts. Toutefois, les co-localisations ne sont pas par défaut des informations que l'on peut représenter spatialement (ce ne sont que des ensembles de caractéristiques booléennes). De plus, les instances participant à chaque co-localisation sont trop nombreuses pour être affichées. Face à ces problèmes, notre idée est de résumer l'information spatiale des instances participant aux co-localisations dans une "couche co-localisations" du SIG.

Le principe de notre approche est de représenter une co-localisation par un ensemble de nouveaux objets spatiaux (générés et stockés dans une couche spécifique du SIG), liés par des traits et positionnés sur la carte. Les liens entre les noeuds représentent la relation de voisinnage entre les caractéristiques. Plus précisément, chaque caractéristique f d'une co-localisation $\mathcal{C}$ est représentée par le centroïde des objets participant à $\mathcal{C}$ et ayant la caractéristique f. En d'autres termes, étant donné $\{o'_1, o'_2, ..., o'_k\}$ un ensemble d'objets spatiaux représentant la co-localisation $C = \{f_1, f_2, ..., f_k\}$, nous avons :

$$o'_i = (x'_i, y'_i), \quad tel\ que \quad x'_i = \frac{\sum_{\forall o=(x,y)\in\Omega_{f_i,C}} x}{|\Omega_{f_i,C}|}, y'_i = \frac{\sum_{\forall o=(x,y)\in\Omega_{f_i,C}} y}{|\Omega_{f_i,C}|} \ et\ \Theta(o'_i) = f_i$$
$$avec\ \Omega_{f_i,C} = \{o \in I | I \in TI_{\mathcal{R}}(\mathcal{O}, \mathcal{C}) \text{ et } \Theta(o) = f_i\}$$

Notre système de visualisation intègre aussi des aspects thématiques en colorant chaque caractéristique de la couleur de son thème. De la même manière, l'importance d'une co-localisation est représentée par la couleur de ses liens. Plus une co-localisation aura un index de participation élevé, plus la couleur de ses liens sera intense.

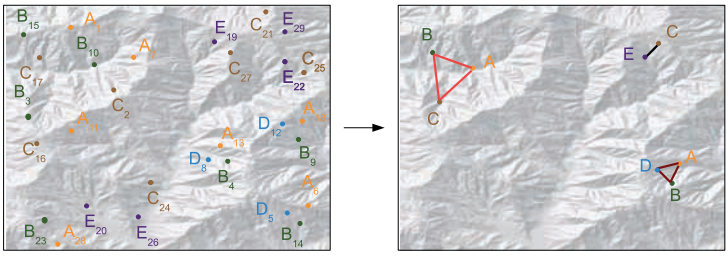

FIG. 2 – *Visualisation cartographique de trois co-localisations*

La figure 2 illustre la visualisation des co-localisations $\{A, B, C\}$, $\{A, B, D\}$ et $\{E, C\}$ (figure de droite) en fonction de leurs instances (figure de gauche). Chaque couleur associée à

une caractéristique correspond à un thème. La couleur des liens de $\{A, B, C\}$ est plus foncée que celle de $\{A, B, D\}$, puisque l'index de participation de $\{A, B, D\}$ est plus grand que celui de $\{A, B, C\}$.

Avantages de notre proposition. Premièrement, nous obtenons une visualisation cartographique des co-localisations totalement intégrée au SIG, répondant ainsi aux besoins et aux pratiques des experts. Les données originelles ne sont pas affectées par notre approche, seul une couche supplémentaire est ajoutée au SIG. De plus, ce système permet de tirer avantage des fonctionnalités offerte par le SIG tel que le zoom. Par exemple, l'utilisateur peut zoomer sur la carte de façon à avoir soit une vision globale de toutes les co-localisations (figure 3 au centre), ou une vue détaillée d'une ou plusieurs co-localisations (figure 3 à droite).

Deuxièmement, cette représentation donne des informations supplémentaires sur les co-localisations. Elle permet notamment de visualiser la localisation globale des objets participant à la co-localisation, i.e. de savoir dans quelle partie de la zone d'étude sont généralement situés ces objets. Par exemple, dans la figure 2 à gauche, la majeure partie des instances participant à la co-localisation $\{A, B, C\}$ sont situées au nord-ouest de la carte. Or, cette particularité est facilement observable en visualisant la couche co-localisation (figure 2 à droite) car la co-localisation est située au nord-ouest de la carte. Notre approche permet aussi de visualiser la distance moyenne entre les objets instanciant la co-localisation. Par exemple, la figure 2 montre que les instances de la co-localisation $\{A, B, D\}$ sont généralement plus proches que celles de la co-localisation $\{A, B, C\}$. De la même manière, l'orientation globale entre les objets instanciant la co-localisation peut être visualisée par notre solution. Par exemple, sur la figure 2), les instances de la co-localisation $\{A, B, D\}$ ont la configuration suivante : les objets ayant la caractéristique B sont en dessous de ceux ayant les caractéristiques A et D, et les objets ayant la caractéristique D sont généralement à gauche de ceux ayant la caractéristique A. De plus, les experts peuvent facilement visualiser l'importance d'une co-localisaion et les thèmes considérés grâce au système de couleurs.

Pour finir, cette approche de visualisation n'implique pas de surcoût supplémentaire, puisque la couche co-localisation est construite pendant l'exécution de l'algorithme d'extraction à partir d'informations déjà utilisées par celui-ci.

Principale limite de notre approche et solution. Toutefois, cette approche de visualisation peut encore poser des problèmes d'interprétation lorsque la co-localisation est située au milieu de la carte. En effet, en pratique, les instances d'une telle co-localisation peuvent être situées au milieu de la carte ou uniformément distribuées sur toute la carte. Ce problème est lié à l'utilisation de centroïdes, calculés à partir de la *moyenne* des coordonnées des objets, pour représenter chaque caractéristique d'une co-localisation.

Une solution pour résoudre ce problème est d'utiliser le clustering de façon à avoir un meilleur regroupement des objets. Chaque co-localisation serait ainsi représentée par plusieurs ensembles de nouveaux objets spatiaux (au lieu d'un seul actuellement). Cette approche est en cours d'intégration dans notre prototype.

5 Application

Les propositions décrites dans ce papier ont été intégrées à un prototype de découverte de co-localisations intéressantes dans un SIG. Ce prototype s'appuie sur l'outil de fouille de

données iZi, proposé dans Flouvat et al. (2009), permettant de résoudre les problèmes de découverte de motifs intéressants tels que définis dans Mannila et Toivonen (1997).

Nous avons utilisé notre prototype pour étudier l'érosion des sols d'un bassin versant montagneux d'environ 9 km^2. Il présente des manifestations de l'érosion naturelle ainsi que des stigmates liés à une activité minière. Lors de cette étude, trois couches thématiques ont été sélectionnées par l'expert : la couche "érosion du sol" (6 caractéristiques), la couche "nature du terrain" (13 caractéristiques) et la couche "végétation" (13 caractéristiques). Les objets d'études étaient des zones géographiques étiquetées par des caractéristiques. La relation spatiale étudiée était une relation de voisinage basée sur une distance maximale entre les centroïdes des zones. Le tableau 1 présente le nombre de co-localisations découvertes pour différentes distances maximum définissant la relation voisinnage et différents seuils minimum d'index de participation pour les co-localisations. Le tableau 2 présente le nombre de co-localisations extraites en utilisant des contraintes définies par un expert géologue. Comme le montre cet exemple, le nombre de motifs découverts peut fortement diminuer en fonction des contraintes, améliorant ainsi les performances, l'interprétation et la pertinence des résultats pour les experts.

	distance 50		distance 100			distance 200				distance 300				
Seuil	taille 2	taille 3	taille 2	taille 3	taille 4	taille 2	taille 3	taille 4	taille 5	taille 2	taille 3	taille 4	taille 5	taille 6
0.1	26	4	66	19	4	116	95	27	4	139	207	96	16	1
0.3	9	2	18	4	0	64	15	2	0	94	53	2	0	0
0.5	8	0	14	3	0	32	6	0	0	59	14	1	0	0
0.7	0	0	7	1	0	14	3	0	0	29	3	0	0	0

TAB. 1 – *Nombre de co-localisations pour l'étude des zones*

	Seuil	Sans contrainte	Maquis ligno-herbacé	Maquis ligno-herbacé & Erosion	Maquis ligno-herbacé & Erosion & Zone
Distance 200	0.1	242	178	206	206
	0.3	81	67	77	77
Distance 300	0.1	459	308	391	391
	0.3	149	107	130	140

TAB. 2 – *Nombre de co-localisations en fonction de contraintes expertes*

La figure 3 illustre la visualisation des co-localisations dans la zone étudiée. La capture d'écran de gauche représente l'ensemble des objets spatiaux toutes couches réunies, celle du centre représente les co-localisations correspondantes, et celle de droite représente un zoom sur une zone donnée.

Nos résultats ont été étudiés et validés par un géologue spécialiste de l'érosion des sols. La connaissance découverte met en avant des corrélations connues sur l'érosion dans cette zone. Elle montre notamment les relations entre les pistes sensibles, les zones minières, l'érosion en rivière et une végétation éparse. Elle souligne très nettement la dégradation du milieu aux alentours des zones où les sols ont été décapés par l'homme.

6 Conclusion et perspectives

Nous nous sommes intéressés dans cet article à la découverte de co-localisations adaptée aux besoins des experts. Dans cet objectif, nous avons proposé une nouvelle approche

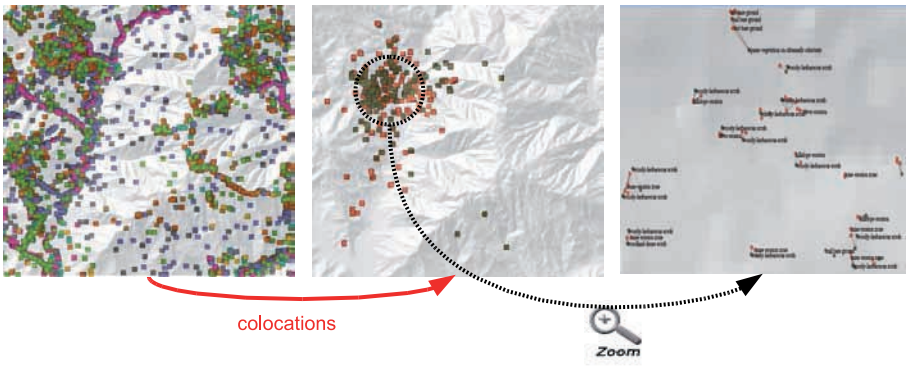

FIG. 3 – *Visualisation des co-localisations pour les données érosion (distance : 200m ; seuil : 0.2)*

permettant d'intégrer les connaissances du domaine dans le processus de découverte des co-localisations, ainsi qu'une nouvelle représentation visuelle de ces motifs. La connaissance des experts a été intégrée dans l'algorithme d'extraction par l'intermédiaire de contraintes thématiques et spatiales. Ces contraintes du domaine permettent de fournir des résultats plus pertinents, tout en améliorant les performances de l'algorithme. D'un point de vue théorique, cette intégration des contraintes dans l'algorithme d'extraction a été rendue possible grâce à l'utilisation d'un cadre théorique existant. Par ailleurs, une nouvelle approche de visualisation cartographique des co-localisations dans un SIG a été développée. Le système proposé permet une représentation simple, concise et intuitive des co-localisations, tout en prenant en considération la nature spatiale des objets sous-jacents et les pratiques des experts. Cette représentation fournit également des informations supplémentaires aux experts, telles que la position moyenne des objets vérifiant la co-localisation, la distance moyenne entre objets ou leur orientation. Ces propositions ont été appliquées à l'étude de l'érosion des sols et validées par un expert du domaine.

Ce travail a plusieurs perspectives. Tout d'abord, la méthode de visualisation pourrait être améliorée pour traiter le cas où les motifs apparaissent au milieu de la carte. Une solution serait d'utiliser un algorithme de clustering afin d'avoir un meilleur regroupement des objets. Cette solution est en cours d'intégration dans notre prototype. Dans certains cas, l'interprétation des résultats par les experts peut être difficile en raison du grand nombre de co-localisations extraites. Face à ce problème, une perspective intéressante serait de proposer une représentation condensée des co-localisations qui serait sans perte d'information. Une autre perspective serait d'améliorer les performances de l'extraction en proposant un nouvel algorithme ou de nouvelles structures de données adaptées aux données spatiales. Pour finir, nous souhaiterions tester notre prototype sur d'autres données et d'autres applications.

Remerciements. Les auteurs souhaitent remercier Isabelle Rouet, géologue et experte en érosion des sols, pour avoir fourni les données et pour avoir validé nos résultats.

Références

Andrienko, G. L. et N. V. Andrienko (1999). Knowledge-based visualization to support spatial data mining. In *IDA*, pp. 149–160.

Bogorny, V., J. Valiati, S. Camargo, P. Engel, B. Kuijpers, et L. O. Alvares (2006). Mining maximal generalized frequent geographic patterns with knowledge constraints. In *IEEE International Conference on Data Mining*, Los Alamitos, CA, USA, pp. 813–817. IEEE Computer Society.

Brunk, C., J. Kelly, et R. Kohavi (1997). Mineset : An integrated system for data mining. In *KDD*, pp. 135–138.

Cao, L. (2008). Domain driven data mining (d3m). In *ICDM Workshops*, pp. 74–76. IEEE Computer Society.

Flouvat, F., F. De Marchi, et J.-M. Petit (2009). The izi project : easy prototyping of interesting pattern mining algorithms. In *Advanced Techniques for Data Mining and Knowledge Discovery*, LNCS, pp. 1–15. Springer-Verlag.

Huang, Y., S. Shekhar, et H. Xiong (2004). Discovering colocation patterns from spatial data sets : A general approach. *IEEE Trans. Knowl. Data Eng. 16*(12), 1472–1485.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Trans. Vis. Comput. Graph. 8*(1), 1–8.

Koperski, K. et J. Han (1995). Discovery of spatial association rules in geographic information databases. In M. J. Egenhofer et J. R. Herring (Eds.), *SSD*, Volume 951 of *Lecture Notes in Computer Science*, pp. 47–66. Springer.

Leung, C. K.-S., P. Irani, et C. L. Carmichael (2008). Wifisviz : Effective visualization of frequent itemsets. In *ICDM*, pp. 875–880. IEEE Computer Society.

Mannila, H. et H. Toivonen (1997). Levelwise search and borders of theories in knowledge discovery. *Data Min. Knowl. Discov. 1*(3), 241–258.

Shekhar, S. et Y. Huang (2001). Discovering spatial co-location patterns : A summary of results. In C. S. Jensen, M. Schneider, B. Seeger, et V. J. Tsotras (Eds.), *SSTD*, Volume 2121 of *Lecture Notes in Computer Science*, pp. 236–256. Springer.

Yoo, J. S. et S. Shekhar (2006). A joinless approach for mining spatial colocation patterns. *IEEE Trans. Knowl. Data Eng. 18*(10), 1323–1337.

Summary

One of the classical task in spatial pattern mining is the extraction of interesting colocations in geo-referenced data. Considering a set of boolean spatial features, the goal is to find subsets of features often located together. However, extracted patterns can represent known relationships for domain experts. Moreover, their interpretation is difficult since they are presented in a textual form. To deal with these problems, we propose in this paper an integration of domain constraints in colocation mining and a new representation of the discovered knowledge. These propositions have been integrated in a prototype with a GIS, experimented on real geological dataset, and validated by a domain expert.