Visualisation de motifs spatiaux dans un SIG

Nazha Selmaoui-Folcher*,**, Frédéric Flouvat*, 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 spatiales. L'objectif est de trouver des sous-ensembles de caractéristiques booléennes apparaissant fréquemment dans des objets spatiaux voisins. Toutefois, l'interprétation des motifs extraits est difficile pour les experts du domaine. En effet, les mesures d'intérêt existantes, utilisées pour filtrer les co-localisations intéressantes, posent des problèmes d'interprétation et les résultats sont présentés aux experts sous forme textuelle. Dans ce papier, nous proposons une nouvelle mesure d'intérêt pour les co-localisations, ainsi qu'une nouvelle représentation visuelle de ces motifs. Notre mesure d'intérêt reflète mieux l'importance d'une co-localisation pour les experts, et est totalement intégrée dans le processus d'extraction. L'approche de visualisation proposée est une représentation simple, concise et intuitive des co-localisations, qui prend 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.

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(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 de la découverte de connaissances centrée sur les données et les contraintes techniques, à une découverte de connaissances centrée sur l'expert.

Une des tâches classiques en fouille de données spatiales est l'extraction de co-localisations intéressantes dans des données spatiales. 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 (Koperski et Han, 1995; Shekhar et Huang, 2001; Huang et al., 2004; Yoo et Shekhar, 2006; Bogorny et al., 2006). Toutefois, une des principales limites de ces travaux est l'interprétation des résultats par les experts. Tout d'abord, les mesures d'intérêt définies dans Shekhar et Huang (2001) et Huang et al. (2004) ne sont pas intuitives pour les experts et peuvent mener à des problèmes d'interprétation. Ensuite, les co-localisations sont présentées aux experts sous forme textuelle, ce qui rend difficile leur utilisation et leur interprétation par les experts. De plus, la nature spatiale des objets sous-jacents n'est pas prise en considération par ces représentations.

Pour pallier à ces problèmes, nous proposons une nouvelle mesure d'intérêt pour les co-localisations, ainsi qu'une nouvelle représentation visuelle de ces motifs. Ces propositions s'appuient sur l'extension du concept de co-localisation dans un cadre théorique existant. La nouvelle mesure d'intérêt reflète mieux l'importance des co-localisations pour les experts, et est totalement intégrée dans l'algorithme d'extraction de motifs. 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 section 3 présente la notion de co-localisation (section 3.1) et une généralisation de ce cadre (section 3.2). La section 4 présente nos propositions pour une découverte des co-localisations adaptée aux besoins des experts : une nouvelle mesure d'intérêt (section 4.1) et une nouvelle approche de visualisation des co-localisations (section 4.2). 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 de référence prédéfinie. Leur méthode énumère les voisinages de la caractéristique spatiale étudiée afin de "matérialiser" un ensemble de transactions correspondant aux instances des caractéristiques voisines. 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)

et 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, ici, toutes les caractéristiques et relations de voisinage sont considérées. 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.

2.2 Visualisation des motifs

Un des défis majeurs en fouille de données est la représentation des connaissances découvertes. Ces connaissances doivent être compréhensibles et directement utilisables par les experts (Han et Kamber, 2006).

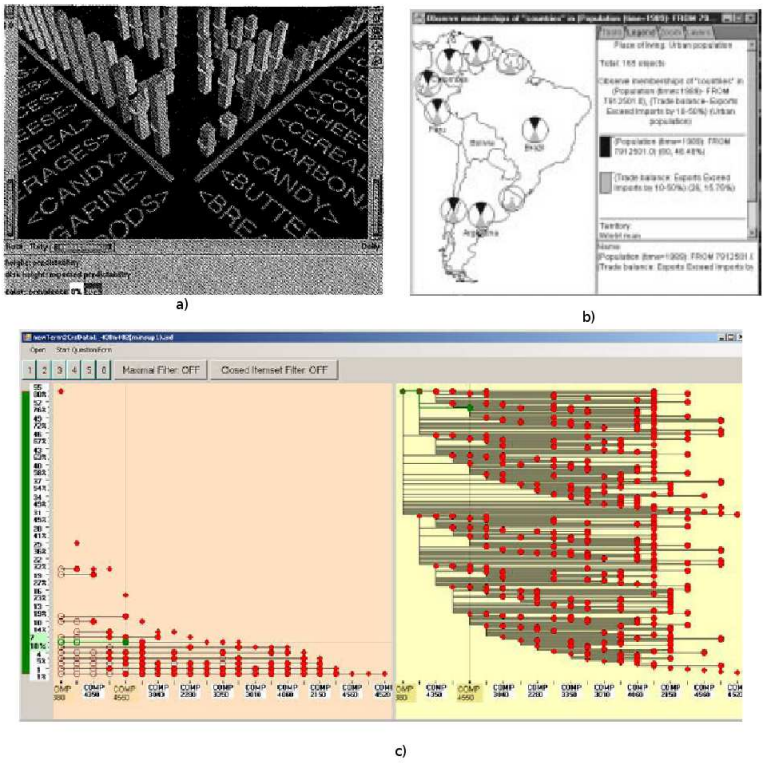

FIG. 1 – *a) Représentation des règles associées aux articles d'un supermarché. b) Visualisation de règles d'association pour l'Amérique du Sud. c) Une capture d'écran de motifs fréquents via WiFIsViz.*

Pour les données classiques, plusieurs systèmes, tels que Brunk et al. (1997) et 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. Différents types de représentations (statistique, arbre, graphique, carte) sont disponibles en fonction du type de résultats à visualiser. Chaque algorithme de fouille de données (e.g. modèle Bayésien, arbre de décision ou règles d'association) est couplé à un outil de visualisation. La figure 1-a montre la visualisation de règles d'association. Les règles sont présentées dans un espace à deux dimensions avec pour axe des abscisses les articles en partie gauche des règles et pour axe des ordonnées les articles en partie droite. A l'intersection, la hauteur de chaque barre représente la confiance de la règle.
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. Le visualiseur fournit différent niveaux de détails pour représenter les motifs, et permet de leur appliquer des contraintes.
On peut également trouver un ensemble d'articles proposant des méthodes de visualisation dans Poulet et Kuntz (2006). Ces méthodes ont toutes des objectifs différents qui vont de la visualisation de données multidimensionnelles en utilisant des cartes auto-organisatrices de Kohenen ou du clustering, à une visualisation des connaissances par des règles d'association en utilisant des représentations hiérarchiques par niveau ou autre. D'autres papier proposent des techniques visuelles de recherche d'information de manière interactive guidée par des experts.

Pour les données spatiales, Andrienko et Andrienko (1999) se sont intéressés à la visualisation des données et des résultats de la fouille de données spatiales. La visualisation de sous-groupes de clusters est naturellement présentée sur une carte en associant des icônes ou des couleurs aux objets spatiaux. 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). La figure1-b illustre une utilisation de ce système de visualisation pour des règles d'association sur des données d'Amérique du Sud.
Une autre méthode de visualisation de règles d'association spatiales a été proposée par Marghoubi et al. (2006). Dans cette méthode, les relations spatiales sont calculées au préalable. Les auteurs s'appuient sur la représentation graphique du treillis de Galois en utilisant la plateforme *Galicia*, qui permet d'afficher des informations sur les noeuds (fermé, son générateur, etc.), pour représenter les règles. Cette représentation est limitée car elle liste les règles d'association en omettant leur contexte spatial. De même, les auteurs Appice et Buono (2005) proposent une méthode de visualisation des règles d'association multi-échelles en utilisant des graphes. Dans ce papier, ils travaillent sur des données à différentes échelles (par exemple quartier, district, région, etc.). Ils extraient des règles d'associations spatiales à chaque niveau, puis les relies de l'échelle la plus petite à la plus grande. Les règles sont ainsi hiérarchisées. Ils utilisent alors un graphe hiérarchisé pour visualiser les règles. Cette méthode permet de visualiser une règle d'association à tous les niveaux d'échelle. Toutefois, aucun aspect spatial n'est pris en compte dans la visualisation, les règles sont uniquement affichées de manière textuelle.

A notre connaissance, les solutions existantes ne prennent pas en considération la nature spatiale des objets étudiés, et utilisent des mesures d'intérêt se focalisant sur des besoins techniques (e.g. l'anti-monotonie du prédicat) au lieu des besoins des experts. Aucune des solutions existantes n'a été conçue pour visualiser des motifs spatiaux de manière simple, concise et intuitive pour les experts.

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) et Huang et al. (2004), puis la généraliser et l'étendre grâce à un cadre théorique existant.

3.1 Présentation du concept de co-localisation

Les co-localisations sont des ensembles de caractéristiques booléennes associées à des objets spatiaux. Ces co-localisations représentent des caractéristiques apparaissant fréquemment dans des objets voisins.

Définition 1: *Soient $\mathcal{F} = \{f_1, f_2, ..., f_k\}$ un ensemble de caractéristiques booléennes et $\mathcal{O} = \{o_1, o_2, ..., o_n\}$ un ensemble d'objets spatiaux. La fonction $\Theta : \mathcal{O} \rightarrow \mathcal{F}$ associe à un objet de $\mathcal{O}$ une unique caractéristique de $\mathcal{F}$ telle que :*

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

La fonction Θ définit formellement l'association entre les objets et les caractéristiques. 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.

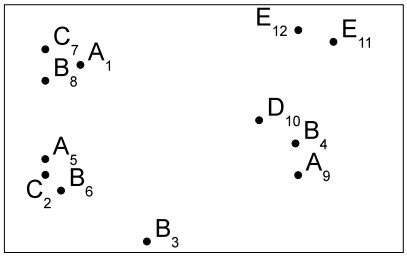

FIG. 2 – *Représentation graphique des objets spatiaux et de leur caractéristique*

Définition 2: *Une **co-localisation** C est un sous-ensemble de caractéristiques de $\mathcal{F}$ tel qu'il existe des ensembles d'objets voisins ayant ces caractéristiques.*

Dans l'exemple de la figure 2, $\{A, B, D\}$ est une co-localisation. Toutefois, toutes les co-localisations ne sont pas intéressantes. En effet, il n'existe qu'un seul ensemble de trois objets

voisins ayant respectivement les caractéristiques $A, B,$ et D. Il convient donc d'introduire un certain nombre de notions afin de déterminer les co-localisations intéressantes.

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

Propriété 1: *Soit $I \subseteq \mathcal{O}$ une instance d'une co-localisation $C \subseteq \mathcal{F}$, par rapport à une relation de voisinage $\mathcal{R}$. On a*
– $\forall o \in I, \Theta(o) = f \; avec \; f \in C$
– $|I| = |C|$
– $\forall o, q \in I, \mathcal{R}(o, q) = vraie$

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. Elle correspond à tous les ensembles d'objets de $\mathcal{O}$ vérifiant la co-localisation C par rapport à la relation de voisinage $\mathcal{R}$. 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 $TI_{\mathcal{R}}(\mathcal{O}, BD) = \{ \{B_4, D_{10}\} \}$

Définition 3: *Soient $C \subseteq \mathcal{F}$ une co-localisation, $\mathcal{O}$ l'ensemble des objets spatiaux et $\mathcal{R}$ une relation de voisinage. La table d'instances de C est :*

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

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. 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 $pr_{\mathcal{R}}(\mathcal{O}, \{A, B, C\}, C) = 1$.

Définition 4: *Soient $C \subseteq \mathcal{F}$ une co-localisation, $\mathcal{O}$ l'ensemble d'objets spatiaux, $\mathcal{R}$ une relation de voisinage et $f \in C$ une des caractéristiques de C. Le ratio de participation de C est égal à :*

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

A partir de ces dernières définitions, les auteurs de Shekhar et Huang (2001) et Huang et al. (2004) ont introduit la notion d'**index de participation** pour évaluer la fréquence et la validité d'une co-localisation dans un jeu de données. Deux définitions de l'index de participation ont été proposées :

– dans (Shekhar et Huang, 2001), l'index de participation noté $pi1$, représente la probabilité d'avoir la co-localisation C parmi tous les objets ayant une caractéristique de C (en supposant que les caractéristiques sont indépendantes).

$$pi1_{\mathcal{R}}(\mathcal{O}, C) = \prod_{\forall f \in C} pr_{\mathcal{R}}(\mathcal{O}, C, f)$$

– dans (Huang et al., 2004), l'index de participation noté $pi2$, représente la probabilité minimale d'avoir un objet ayant une caractéristique de la co-localisation C parmi l'ensemble des objets ayant cette même caractéristique.

$$pi2_{\mathcal{R}}(\mathcal{O}, C) = \min_{\forall f \in C}(pr_{\mathcal{R}}(\mathcal{O}, C, f))$$

Quelque soit la définition de l'index de participation choisie, le problème à résoudre est le suivant :

Problème de l'extraction des co-localisations
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 de proposer une première représentation condensée des co-localisations.

3.2.1 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 de 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). Lorsque le prédicat est anti-monotone, l'ensemble $Th(\mathcal{L}, d, Q)$ est dit "fermé par le bas". Dans ce cas, il peut être représenté de façon équivalente par sa bordure positive ou négative :

– *sa bordure positive*, notée $\mathcal{B}d^{+}(Th(\mathcal{L}, d, Q))$, est composée des motifs intéressants les plus spécialisés par rapport à la relation d'ordre.
– *sa bordure négative*, notée $\mathcal{B}d^{-}(Th(\mathcal{L}, d, Q))$, est composée des motifs non intéressants les plus généraux par rapport à la relation d'ordre.

Si le prédicat est monotone, l'ensemble $Th(\mathcal{L}, d, Q)$ est dit "fermé par le haut", et la notion de bordures existe toujours. La bordure positive est alors composée des motifs intéressants les

plus généraux, et la bordure négative des motifs non intéressants les plus spécialisés. L'intérêt de chacune de ces bordures est de permettre de représenter la théorie (les motifs intéressants). Ainsi, il est possible de déterminer si un motif est intéressant ou non en étudiant uniquement les motifs d'une des deux bordures, et ceci sans avoir à accéder aux données.

3.2.2 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 supérieur (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.

De plus, tout algorithme défini pour un problème d'extraction de motifs intéressants peut être utilisé pour la découverte de co-localisations intéressantes. La majeure partie des algorithmes d'extraction de motifs fréquents tels que Agrawal et Srikant (1994); Zaki et al. (1997); Gouda et Zaki (2001); Gunopulos et al. (2003) peuvent donc être utilisés pour extraire les co-localisations.

Utilisation des bordures pour résumer les co-localisations intéressantes Pour finir, la notion de bordure peut être étendue aux co-localisations. La bordure positive des co-localisations intéressantes est l'ensemble des co-localisations intéressantes maximales par rapport à l'inclusion. La bordure négative est constituée de l'ensemble des co-localisations non intéressantes minimales par rapport à l'inclusion.

A partir de la seule bordure positive, on peut décider si une co-localisation quelconque C est intéressante ou non – et ce sans accès aux données. En effet, s'il existe une co-localisation C' de la bordure positive telle que $C \subseteq C'$ alors C est intéressante, autrement C n'est pas intéressant. La bordure positive constitue ainsi une première représentation condensée (c.a.d. un résumé) des co-localisations intéressantes. Notons néanmoins que lorsqu'une mesure est associée aux co-localisations intéressantes (e.g. l'index de participation), les bordures ne suffisent pas pour "recalculer" cette mesure.

4 Vers une visualisation des co-localisations adaptée aux besoins des experts

Nous proposons dans cette section une nouvelle mesure d'intérêt et une nouvelle approche de visualisation facilitant l'interprétation et l'utilisation par les experts des motifs spatiaux découverts.

4.1 Proposition d'une nouvelle mesure d'intérêt pour les experts

Comme nous l'avons vu en section 3.1, deux mesures d'intérêt ont été définies dans Shekhar et Huang (2001) et Huang et al. (2004) pour sélectionner les co-localisations intéressantes. Toutefois, ces deux mesures reflètent mal l'importance d'une co-localisation dans les données. Nous proposons ici une nouvelle mesure d'intérêt, appelé **ratio de co-localisation**, dont l'objectif est de faciliter et d'affiner l'interprétation des résultats par les experts.

La première mesure $pi1_{\mathcal{R}}(\mathcal{O}, C) = \prod_{\forall f \in C} pr_{\mathcal{R}}(\mathcal{O}, C, f)$ représente la probabilité d'avoir la co-localisation C parmi tous les objets ayant une caractéristique de C, en supposant que les caractéristiques soient indépendantes. Or, cette hypothèse va à l'encontre de notre objectif qui est de découvrir les relations spatiales entre les caractéristiques. Cette mesure n'est donc pas adaptée pour représenter l'importance d'une co-localisation. L'exemple de la figure 3 le démontre : l'index de participation de $\{A, B\}$ est égal à 1/4 alors que la moitié des objets ayant pour caractéristique A ou B participent à cette co-localisation.

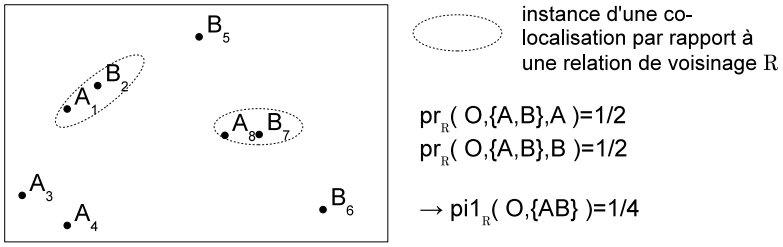

FIG. 3 – *Exemple de problème d'interprétation avec* $pi1$

La deuxième mesure $pi2_{\mathcal{R}}(\mathcal{O}, C) = \min_{\forall f \in C}(pr_{\mathcal{R}}(\mathcal{O}, C, f))$ est plus pertinente que la première, mais rend difficile l'interprétation des résultats par les experts. En effet, l'importance d'une co-localisation C dépend uniquement de la caractéristique de C dont la probabilité de participer à la co-localisation est la plus faible. Cette mesure reflète parfois mal l'importance globale de la co-localisation. Par exemple dans la figure 4 de gauche, la co-localisation $\{A, B\}$ a le même index de participation bien que celle-ci soit trois fois plus présente dans la figure 4 de droite. Cette mesure ne permet donc pas de différencier ces deux cas, pourtant très différents pour les experts.

Pour pallier à ces problèmes, nous avons défini conjointement avec un expert, une nouvelle mesure appelée **ratio de co-localisation** (notée $cr_{\mathcal{R}}(\mathcal{O}, C)$) afin d'améliorer l'interprétation des résultats. L'indicateur proposé est égal au ratio du nombre d'instances vérifiant la co-localisation étudiée sur le nombre minimal d'instances qu'il y aurait si la co-localisation

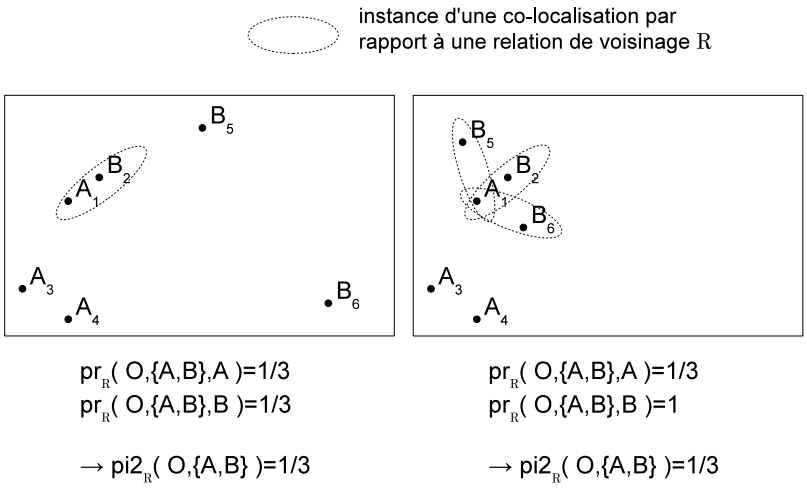

FIG. 4 – *Exemple de problème d'interprétation avec* $pi2$

était toujours vérifiée. Par exemple sur la figure 4, le nombre d'instances de la co-localisation $\{A, B\}$ est égal à trois et le nombre minimal d'instances qu'il y aurait si celle-ci était toujours vérifiée est égal à cinq. En effet, il y a trois instances qui vérifient réellement cette co-localisation, plus deux objets ayant la caractéristique A mais qui ne la vérifient pas. Or, si cette co-localisation était toujours vraie, ces deux objets seraient nécessairement à proximité de deux objets ayant la caractéristique B, et on aurait ainsi deux nouvelles instances. Le ratio de co-localisation est donc égal à $3/5$, ce qui représente mieux l'importance de la co-localisation.

Définition 5: *Soient $C \subseteq \mathcal{F}$ une co-localisation, $\mathcal{O}$ l'ensemble des objets spatiaux étudiés, $\mathcal{R}$ une relation spatiale booléenne.*

$$cr_{\mathcal{R}}(\mathcal{O}, C) = \frac{|TI_{\mathcal{R}}(\mathcal{O}, C)|}{|TI_{\mathcal{R}}(\mathcal{O}, C)| + \max\limits_{\forall f \in C} |\{o \in TI_{\mathcal{R}}(\mathcal{O}, f) \mid o \notin I, \forall I \in TI_{\mathcal{R}}(\mathcal{O}, C)\}|}$$

La figure 5 montre sur les mêmes exemples que le ratio de co-localisation n'a pas les inconvénients des deux mesures d'index de participation. Notre mesure donne une information plus précise sur la représentativité de la co-localisation parmi les objets spatiaux. De plus, sa définition intuitive facilite l'interprétation des résultats par les experts. En effet, outre les problèmes précédents, la définition même des mesures d'intérêt existantes est difficilement compréhensible pour les experts. Un autre avantage de cette mesure est qu'elle peut être calculée sans surcoût durant l'extraction des motifs, puisqu'elle utilise des informations déjà utilisées pour calculer l'index de participation.

Notons néanmoins que, contrairement aux indexes de participation, la propriété $cr_{\mathcal{R}}(\mathcal{O}, C) \geq \sigma$ n'est pas anti-monotone (ni monotone). Elle ne peut donc pas être utilisée pour élaguer l'espace de recherche dans l'algorithme de fouille de données. L'approche proposé dans ce papier utilisera donc l'index de participation $pi2$ pour la sélection des co-

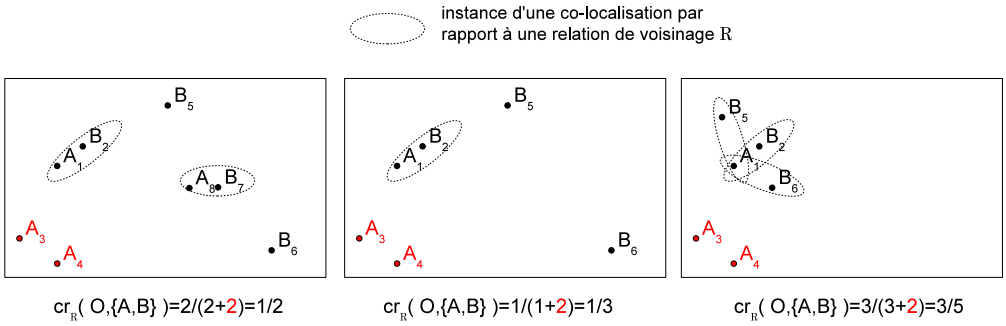

FIG. 5 – *Exemple de mesures obtenues avec le ratio de co-localisation*

localisations intéressantes, et le ratio de co-localisation sera utilisé comme indicateur pour l'expert dans l'interface de visualisation.

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

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, 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. Notre idée est donc de résumer l'information spatiale des instances participant aux co-localisations dans une "couche co-localisations" du SIG (figure 6).

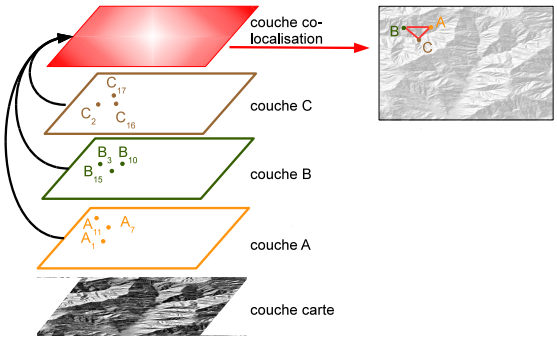

FIG. 6 – *Principe de notre approche*

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. Autrement dit, une co-localisation est représentée par une clique, où les points représentent ses objets et les traits représentent la relation de voisinage. Cette représentation sous forme de clique a été en premier lieu utilisée dans les exemples de Shekhar et Huang (2001) pour représenter les instances des co-localisations. Cette représentation a aussi été choisie pour notre système de visualisation car elle était facilement compréhensible pour les experts. D'autres représentations ont été étudiées, tel que des triangles pleins, toutefois elles n'étaient pas aussi claires et simples au regard des experts.

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. En effet, classiquement, les SIG regroupent les caractéristiques en couches thématiques et associent à ces couches des couleurs. Par exemple, les caractéristiques d'une couche thématique "Etat du sol" (p.ex. "sol non nu", "sol nu sur piste", "érosion de versant" ou "érosion sur mine") peuvent être associés à la couleur rouge. Ainsi, il est facile pour les experts de visualiser quels sont les thèmes abordés par les informations affichées à l'écran. Dans notre contexte, les couleurs permettent donc de visualiser rapidement les thèmes des caractéristiques composant les co-localisations, et donc par extension les corrélations entre thèmes. Supposons par exemple que les thèmes "Etat du sol" et "Végétation" sont associés aux couleurs rouge et verte. La présence d'un grand nombre de co-localisations ayant des caractéristiques rouges et vertes mettra en avant le lien fort entre l'état du sol et le type de végétation. Autrement dit, état du sol et végétation sont fortement corrélés. Pour finir, notons qu'en pratique le nombre de thèmes est limité (en général quelques dizaines) bien que chacun des thèmes puisse comporter un nombre important de caractéristiques. La question du nombre de couleurs et de leur répartition entre les différents thèmes ne se pose donc pas.

De la même manière, l'importance d'une co-localisation est représentée par la couleur de ses liens. En effet, l'expert a trouvé l'utilisation des couleurs pour les thèmes particulièrement intéressante. Par conséquent, nous avons aussi opté pour un système à base de couleurs pour représenter l'importance d'une co-localisation. La mesure d'intérêt étant continue et les couleurs représentant les différents thèmes, nous avons choisi une représentation exploitant l'intensité lumineuse. L'avantage de cette approche est de mettre en avant simplement les co-localisations les plus importantes car l'attention de l'utilisateur est naturellement attirée en premier lieu par ce qui est plus lumineux. Par conséquent, plus une co-localisation aura un ratio de co-localisation élevé, plus la couleur de ses liens sera intense.

La figure 7 illustre la visualisation des co-localisations $\{A, B, C\}$ et $\{A, B, D\}$ (figure de gauche) en fonction de leurs instances (figure de droite). Chaque couleur associée à une caractéristique correspond à un thème. La couleur des liens de $\{A, B, C\}$ est plus foncée que

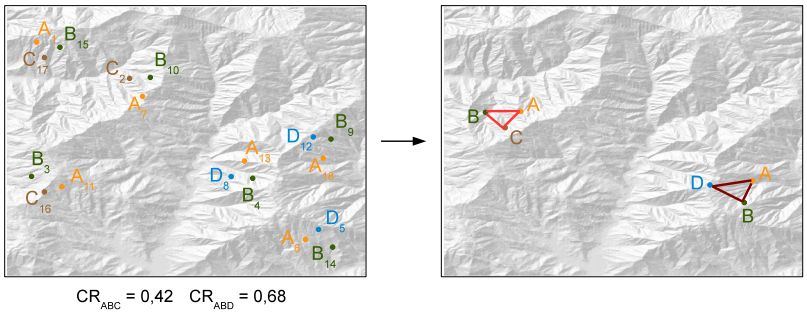

FIG. 7 – *Visualisation de deux co-localisations de taille 3*

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. Cette approche de visualisation a les avantages suivants :
- elle est totalement intégrée au SIG,
- elle fournit des informations spatiales sur les objets liés aux co-localisations,
- elle est intégrée à l'algorithme d'extraction (elle n'est pas un post-traitement).

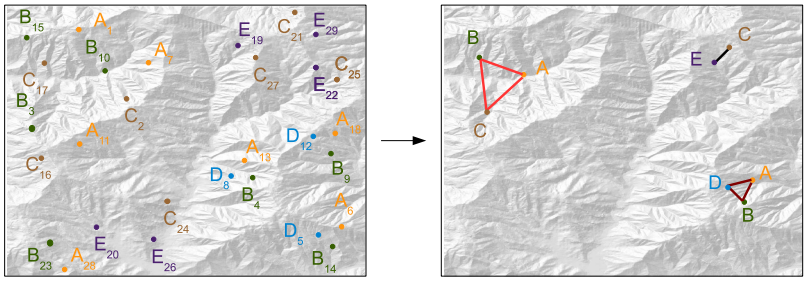

FIG. 8 – *Visualisation des informations spatiales des co-localisations : localisation, distance et orientation*

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 12 au centre), ou une vue détaillée d'une ou plusieurs co-localisations (figure 12 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 8 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 8 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 8 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 8, 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 traitements supplémentaires, puisque la couche co-localisation est construite pendant l'exécution de l'algorithme d'extraction à partir d'informations déjà utilisées par celui-ci.

Principales limites de notre approche et solutions. Toutefois, cette approche de visualisation présente quelques limites. Tout d'abord, l'interprétation peut encore être difficile si beaucoup de co-localisations sont générées. La fonction de zoom du SIG résout partiellement ce problème, mais dans certains cas cela peut ne pas être suffisant. Pour résoudre ce problème, l'utilisateur peut choisir d'extraire une représentation condensée des co-localisations intéressantes, i.e. un sous-ensemble de co-localisations représentant les solutions. Ainsi, notre système propose aussi l'extraction et la visualisation de la bordure positive à la place de toutes les co-localisations intéressantes (cf section 3.2.2). Les expérimentations présentées en section 5.2 mettront en avant l'intérêt de cette bordure pour diminuer le nombre d'objets présentés à l'expert.

D'autre part, notre approche de visualisation peut aussi 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 (figure 9). 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 (figure 10). 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.

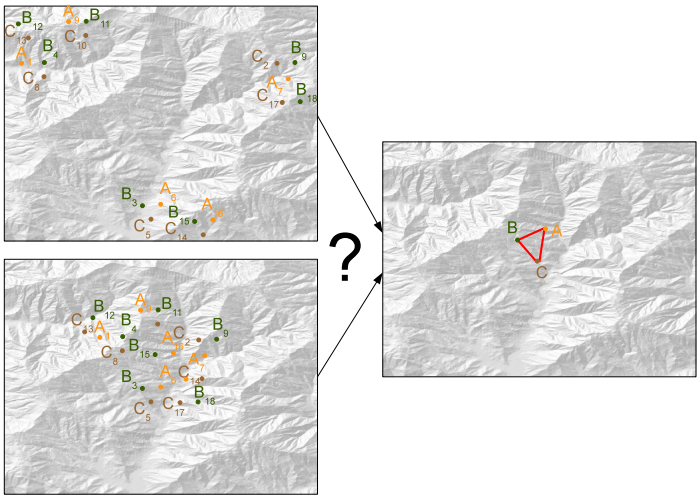

FIG. 9 – *Exemple de problème d'interprétation lié aux centroïdes*

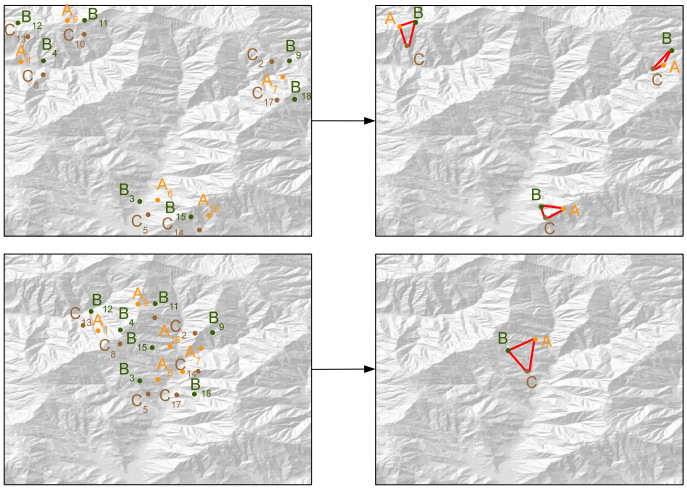

FIG. 10 – *Exemple de visualisation utilisant le clustering*

5 Application

5.1 Prototype

Les propositions décrites dans ce papier ont été intégrées à un prototype de découverte de co-localisations intéressantes dans un SIG (voir figure 11). 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 dans le cadre théorique de Mannila et Toivonen (1997). L'intérêt de cet outil est de fournir un ensemble d'algorithmes génériques (e.g. $Apriori$ de Agrawal et Srikant (1994) et ABS de De Marchi et al. (2005)) pour ce type de problèmes.

Cet outil a été complété par deux composants. Le premier permet d'extraire des co-localisations intéressantes tout en calculant le ratio de co-localisation. Le deuxième permet de stocker les solutions (tous les motifs et/ou la bordure positive) dans une table de la base de données géographiques $PostGis$. Cette table est ensuite utilisée en tant que "couche résultat" lors de la visualisation des résultat par le logiciel $uDig$.

Notre prototype prend trois paramètres en entrée : la relation spatiale à étudier, le seuil minimum utilisé pour sélectionner les co-localisations intéressantes, et les paramètres de la base de données.

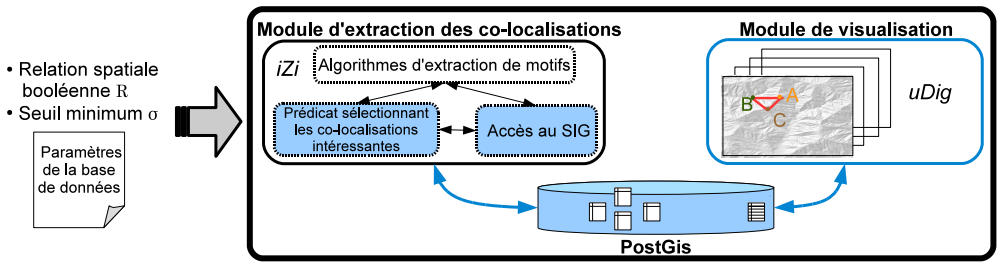

FIG. 11 – *Architecture du prototype*

5.2 Expérimentations

5.2.1 Protocole expérimental

Description des données L'application a été réalisée 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 7700 objets associés à des caractéristiques regroupés en trois couches. Les trois couches thématiques considérées comme importantes pour les experts pour l'étude de l'érosion (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 tel que décrit par 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 telles que "alluvions actuelles et récentes", "harzburgites", "serpentinites" ou "latérites indifférenciées sur péridotites". Elle est issue du SIG géologique à l'échelle du 1/50 000 de DIMENC/SGNC et 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 tels que "savane", "maquis ligno-herbacé", "forêt sur substrat volcano-sédimentaire", ou "végétation éparse sur substrat ultramafique". Elle a été obtenue par télédétection par DTSI/SGT (2008).

Devant le grand nombre de polygones disponibles pour la zone d'étude et après avoir constaté un temps de calcul important, l'étude a dans un premier temps été focalisée sur les centroïdes des polygones. Les objets d'études étaient donc des zones géographiques étiquetées chacune par une caractéristique et représentée par son centroïde. La relation spatiale étudiée était une relation de voisinage basée sur une distance maximale entre les centroïdes des zones.

Description du protocole d'évaluation de l'approche Afin de valider notre approche, nous avons fait analyser les résultats de notre système de visualisation par une géologue spécialiste de l'érosion des sols de la zone d'étude.

La quantification de l'aide apportée par un système de visualisation est difficile comme l'indique des travaux tel que ceux de Sebrechts et al. (1999); Marghescu et al. (2004); Zhao et al. (2005). Afin de valider leur approche, un certain nombre de travaux mettent en place un protocole de test impliquant des "utilisateurs lambda" (p.ex. une dizaine d'étudiants). Un formulaire d'évaluation leur est ensuite remis afin de quantifier l'intérêt du système de visualisation.

Dans notre cas, la mise en place d'un protocole de validation à une telle échelle est difficile en raison de la spécificité de l'application étudiée. En effet, pour valider l'apport de notre approche en terme d'informations et d'utilisation, nous ne pouvons faire appel qu'à des experts de l'érosion des sols en milieu sub-tropical. Les phénomènes observés sont très spécifiques et requiert un niveau d'expertise élevé pour pouvoir être pleinement analysés. Un "utilisateur lambda" pourrait difficilement dire si l'information extraite correspond à une "réalité terrain". Par exemple, un étudiant pourrait difficilement valider le fait que les "pistes sensibles" sont liées à la "nature du sol latéritique" dans la partie nord-ouest de la zone d'étude.

L'évaluation des résultats par l'expert a été faite en deux temps :
- Tout d'abord, l'expert a analysé les co-localisation extraites afin d'évaluer si elles correspondaient bien à une réalité géologique. Plusieurs distances ont été testées (50, 100, 200 et 300 m) avec différents seuils d'index de participation (0.1, 0.3, 0.5, 0.7, et 0.9).
- Ensuite, l'expert a évalué l'information spatiale apportée par notre système, ainsi que son interface visuelle, par rapport aux principales alternatives pour visualiser des co-localisations. Par ailleurs, nous avons aussi étudié le nombre d'informations affichées à l'expert afin d'évaluer notre approche visant à "résumer" les co-localisations.

Plus précisément, nous avons présenté à l'expert trois alternatives de visualisation des co-localisations. La première présente une information textuelle aux experts, i.e. une liste de co-localisations intéressantes avec leur mesure. Elle correspond à l'approche utilisée classique-

ment pour visualiser des motifs (spatiaux ou pas). La deuxième propose à l'expert de sélectionner un motif intéressant (un à la fois) à partir d'une liste textuelle, puis affiche sur une carte toutes les instances de la co-localisation choisie. Elle correspond à l'approche classique utilisée pour visualiser des résultats de fouille de données spatiales (c'est notamment l'approche utilisée par Andrienko et al.). La troisième affiche sur une carte l'ensemble des co-localisations représentées par des cliques et placées sur la carte en fonction de la position moyenne de leurs instances. Elle correspond à l'approche proposée dans cet article.

5.2.2 Présentation des résultats et analyse par l'expert

Evaluation par l'expert de l'information apportée par notre approche Le tableau 1 présente le nombre de co-localisations découvertes pour différentes distances et seuils minimum d'index de participation.

	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 la zone éudiée*

Ces résultats ont été analysés par l'expert. Ses conclusions sont les suivantes : *" Les co-localisations intra-thème (c.a.d. entre des caractéristiques d'un même thème) sont essentiellement liées au thème "Etat du sol", avec un ratio de co-localisation 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 co-localisations 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 par exemple). Aucune co-localisation 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 co-localisations dans cette dernière couche qui peut présenter des polygones de grande superficie, soit une distance entre centroïdes au-delà des seuils choisis. Les co-localisations inter-thèmes (c.a.d. entre des caractéristiques de thèmes différents) 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."*

Evaluation de notre approche visant à résumer les co-localisations Le tableau 2 présente le nombre de co-localisations dans la bordure positive (si les experts ont choisi de visualiser une

représentation condensées des co-localisations). Ces expérimentations montrent que l'utilisation de la bordure positive pour résumer les co-ocalisations intéressantes, permet de diminuer de manière importante le nombre d'objets affichés à l'écran. Cette diminution est d'autant plus importante que le nombre de co-localisations est élevé. Par exemple, le nombre de solutions affichées à l'écran passe de 459 à 108, soit près de 77% de réduction, pour une distance de 300m et un seuil de participation minimum de 0.1. L'utilisation de ce résumé améliore ainsi les performances, l'interprétation et la pertinence des résultats pour les experts.

	distance 50		distance 100		distance 200		distance 300	
Threshold	coloc Bd+	% reduc	coloc Bd+	% reduc	coloc Bd+	% reduc	coloc Bd+	% reduc
0.1	22	26,67	50	43,82	77	68,18	108	76,47
0.3	5	54,55	12	45,45	47	41,98	69	53,69
0.5	8	0	10	41,18	25	34,21	44	40,54
0.7	0	0	5	37,5	10	41,18	25	21,88
0.9	0	0	1	0	8	0	14	17,65

TAB. 2 – *Nombre de co-localisations dans la bordure positive et pourcentage de réduction par rapport au nombre total de co-localisations*

Evaluation par l'expert de l'apport de notre système de visualisation La figure 12 illustre un exemple de visualisation de co-localisations pour 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.

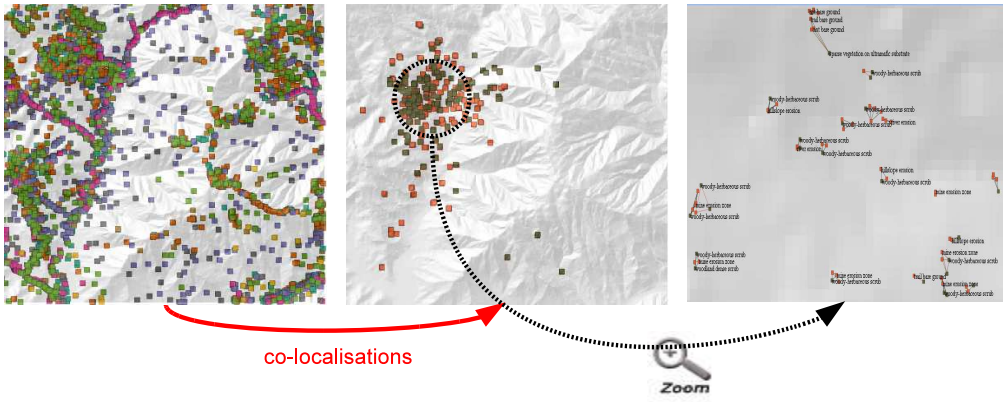

FIG. 12 – *Visualisation des co-localisations pour les données érosion*

Afin d'évaluer les qualités et les limites de notre approche, notre expert a analysé les résultats obtenus avec un même seuil et une même distance (0.3 et 200 m) à partir de trois approches de visualisation différentes.

Dans un premier temps, nous avons fourni à l'expert la liste des co-localisations intéressantes ainsi que leur mesure (approche basique utilisée dans la majeure partie des travaux

en fouille de données). Avant de commencer son étude, l'expert a préparé la liste des co-localisations. Il a notamment recherché les co-localisations les plus importantes et les a colorées. La liste textuelle des co-localisations lui a permis de qualifier et quantifier les relations entre les différentes caractéristiques. Toutefois, une information importante manquait : l'information sur la répartition spatiale des instances de ces co-localisations. Afin de pouvoir répondre à cette question, l'expert a dû étudier les principales co-localisations les une après les autres, et pour chacune d'entre elles, il a dû écrire et exécuter une requête dans le SIG afin de visualiser les objets concernés. L'ensemble de l'étude lui a pris près d'une journée et à nécessité un travail fastidieux.

Dans un second temps, nous avons fourni à l'expert la liste des co-localisations avec pour chacune d'entre elles un "lien" affichant uniquement les objets correspondants sur la carte (approche de visualisation classique en fouille de données spatiales). Ce lien est en réalité une requête SIG stockée permettant de filtrer les instances d'une co-localisation. Ainsi, l'expert n'a plus besoin d'effectuer de requêtes pour chaque co-localisation sélectionnée (comme cela est le cas dans l'approche précédente). L'intérêt de ce système est d'afficher à l'écran uniquement les objets appartenant aux instances d'une co-localisation choisie, donnant ainsi une vision détaillée de sa répartition spatiale. De plus, l'expert a pu utiliser certaines fonctionnalités du SIG, telle que l'utilisation de couleurs pour les différentes caractéristiques, afin de facilité son analyse. Le principal problème de cet approche d'après l'expert est qu'il doit toujours pré-sélectionner une co-localisation avant de pouvoir étudier la répartition spatiale de ses instances.

Pour finir, nous avons proposé à l'expert de visualiser les co-lcalisations par notre système. L'analyse des données par l'expert a montré que notre approche permettait d'avoir une vision globale de la spatialisation des co-localisations. Le système a notamment permis de mettre en avant des phénomènes isolés telles que des associations végétales particulières, spécifiques des fonds de vallées proches de la Côte Ouest de l'île. Toutefois, pour certaines co-localisations, l'utilisation des centroïdes a montrer ses limites. Pour l'expert, l'analyse est malgré tout possible, avec le besoin de revenir à un niveau de détail supérieur (en utilisant par exemple l'approche précédente). L'intérêt principal d'une vue globale est d'identifier rapidement, par exemple, des phénomènes nouveaux, isolés ou au contraire très généraux. Contrairement aux approches précédentes, la sélection de certaines co-localisations ne se fait donc plus uniquement en fonction de la mesure d'intérêt mais aussi en fonction de critères spatiaux, ce qui constitue une réelle amélioration pour les experts.

Pour résumer, d'après l'expert, les compléments apportés par ces résultats préliminaires à l'étude de l'érosion sont significatifs. L'approche de visualisation proposée est complémentaire de celle affichant toutes les instances d'une co-localisation sélectionnée (approche utilisée entre autres dans les travaux de Andrienko et al.). En effet, notre approche fournit une vision globale de la répartition spatiale des co-localisations alors que l'autre approche fournit une vision détaillée pour une co-localisation pré-sélectionnée. Il est notamment possible de caractériser de manière plus pertinente des associations fréquentes, tout en y associant une quantification et spatialisation des relations entre objets difficiles à obtenir par ailleurs de manière globale. 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. L'étude des co-localisations met également en avant les systèmes végétaux qui peuvent être liés à la dégradation du milieu.

6 Conclusion et perspectives

Nous nous sommes intéressés dans cet article à la mise en place d'un système de visualisation des co-localisations adapté aux besoins des experts. Dans cet objectif, nous avons proposé une nouvelle mesure d'intérêt améliorant l'interprétation des co-localisations par les experts ainsi qu'une nouvelle représentation visuelle de ces motifs. La mesure d'intérêt reflète mieux l'importance d'une co-localisation pour les experts, et est calculée lors de l'extraction sans surcoût supplémentaire. Cette mesure a ensuite été utilisée dans un nouveau système de visualisation cartographique des co-localisations dans un SIG. 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 les objets ou leur orientation. Grâce à l'extension du cadre théorique, nous avons également introduit une première représentation condensée des co-localisations, permettant ainsi d'améliorer la visualisation des motifs lorsque ceux-ci sont trop nombreux. Ces propositions ont été appliquées à l'étude de l'érosion des sols et validées par un expert du domaine.

Ce travail offre 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

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules in large databases. In J. B. Bocca, M. Jarke, et C. Zaniolo (Eds.), *VLDB*, pp. 487–499. Morgan Kaufmann.

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

Appice, A. et P. Buono (2005). Analyzing multi-level spatial association rules through a graph-based visualization. In *IEA/AIE*, pp. 448–458.

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.

De Marchi, F., F. Flouvat, et J.-M. Petit (2005). Adaptive strategies for mining the positive border of interesting patterns : Application to inclusion dependencies in databases. In J.-F. Boulicaut, L. De Raedt, et H. Mannila (Eds.), *Constraint-Based Mining and Inductive Databases*, Volume 3848 of *Lecture Notes in Computer Science*, pp. 81–101. Springer.

DIMENC/SGNC et BRGM (2005). Carte géologique de la nouvelle-calédonie au 1/50ă000.

DTSI/SGT (2008). Cartographie de l'occupation du sol de la nouvelle-calédonie au 1/50ă000.

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.

Gouda, K. et M. J. Zaki (2001). Efficiently mining maximal frequent itemsets. In N. Cercone, T. Y. Lin, et X. Wu (Eds.), *ICDM*, pp. 163–170. IEEE Computer Society.

Gunopulos, D., R. Khardon, H. Mannila, S. Saluja, H. Toivonen, et R. S. Sharm (2003). Discovering all most specific sentences. *ACM Trans. Database Syst. 28*(2), 140–174.

Han, J. et M. Kamber (2006). *Data Mining, Second Edition : Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems) (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann.

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.

Marghescu, D., M. Rajanen, et B. Back (2004). Evaluating the quality of use of visual data-mining tools. In *European Conference on IT Evaluation*, pp. 239–250.

Marghoubi, R., A. Boulmakoul, et K. Zeitouni (2006). Utilisation des treillis de galois pour l'extraction et la visualisation des règles d'association spatiales. In *INFORSID*, pp. 703–718.

Poulet, F. et P. Kuntz (2006). Visualisation en Extraction des Connaissances. *RNTI E-7*. 182p.

Rouet, I., D. Gay, M. Allenbach, N. Selmaoui, A.-G. AUSSEIL, M. Mangeas, J. Maura, P. Du-

mas, et D. Lille (2009). Tools for soil erosion mapping and hazard assessment : application to new caledonia, sw pacific. In *International Congress on Modelling and Simulation (MODSIM'09), Cairns, Australia*, pp. 1986–1992.

Sebrechts, M. M., J. Cugini, S. J. Laskowski, J. Vasilakis, et M. S. Miller (1999). Visualization of search results : A comparative evaluation of text, 2d, and 3d interfaces. In *SIGIR*, pp. 3–10. ACM.

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.

Zaki, M. J., S. Parthasarathy, M. Ogihara, et W. Li (1997). New algorithms for fast discovery of association rules. In *KDD*, pp. 283–286.

Zhao, K., B. Liu, T. M. Tirpak, et W. Xiao (2005). A visual data mining framework for convenient identification of useful knowledge. In *ICDM*, pp. 530–537. IEEE Computer Society.

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, the interpretation of the extracted patterns by domain experts is difficult. Indeed, existing interestingness measures can lead to interpretation problems, and solutions are presented in a textual form. To deal with these problems, we propose in this paper a new interestingness measure for colocations and a new representation of the discovered knowledge, based on an existing theoretical framework. Our interestingness measure better reflects the importance of a colocation for the experts, and is totally integrated in the mining process. Our visualization approach is a simple, concise and intuitive representation of the colocations, that takes into consideration the spatial nature of the underlying objects and the experts practice. These propositions have been integrated in a prototype with a GIS, experimented on real geological dataset, and validated by a domain expert.