Appendix B

RÉSUMÉ EN FRANÇAIS

B.1. Introduction

Durant ces dernières années nous avons été témoins de la rapide croissance du nombre, de la capacité et de la dissémination des applications informatiques consacrées à l'obtention, la génération, la manipulation, et le stockage de données dans divers milieux de la vie humaine. Ces actions ont impliqué le recueil d'une grande quantité de données dont l'analyse par des moyens manuels devient chaque jour plus compliquée. Rappelons que dans plusieurs occasions les données “crues” ont besoin d'être analysées et interprétées pour qu'elles soient transformées en informations utiles et profitables. Cette situation a été résolue grâce à des techniques et des outils informatiques qui, de plus en plus nombreux, nous aident pour atteindre ces objectifs. La découverte de connaissances dans bases de données (*KDD*, sigle en langue anglaise) est définie comme l'extraction non banale d'informations implicites, préalablement inconnues et potentiellement utiles à partir de données [16]. Il s'agit d'un processus itératif et interactif qui comprend différentes phases. Le noyau du processus est la phase du data mining (fouille de données), qui peut être conceptualisée comme l'application d'algorithmes d'analyse de données et de fouille de

données qui grâce à des paramètres adéquats produisent et découvrent une énumération particulière de patrons (régularités) sur des données elles-mêmes [12].

Dans ce contexte, l'analyse et l'exploitation de données provenant de phénomènes existant en, sur, et sous la surface de la terre – appelés données spatiales –, ont engendré un nouveau domaine d'investigation appelé fouille de données spatiales, de telle sorte que la fouille de données spatiales s'envisage comme la découverte de connaissances implicites, et préalablement inconnues de données spatiales [16]. Diverses approches pour la fouille de données spatiales ont été développées, dont ne seront examinées que les plus représentatives.

B.1.1 Méthodes basées sur la généralisation

La généralisation a démontré être une des méthodes effectives pour découvrir des connaissances. Elle a été introduite par la communauté d'apprentissage automatique et se base sur l'apprentissage à partir d'exemples. La découverte de connaissances basée sur la généralisation requiert la construction de hiérarchies de concepts (donnés explicitement par l'expert ou générés automatiquement). Dans le cas des bases de données spatiales, on peut trouver deux types de hiérarchies de concepts: (1) Hiérarchies thématiques, par exemple, généraliser des tomates et des bananes comme des fruits, les fruits et légumes comme des aliments d'origine végétale, etc. (2) Hiérarchies spatiales, par exemple, généraliser une série de points spatiaux comme une région ou pays.

Lu et al. [35] étendent une technique d'induction basée sur les attributs, aux bases de données spatiales. Cette technique construit la hiérarchie de généralisation en résumant les

relations entre les données spatiales et non spatiales à un niveau de concept plus élevé. Les auteurs présentent deux algorithmes basés sur la généralisation: (1) Approche de domination de données non spatiales. Cette méthode réalise en première instance une induction basée sur les attributs aux données non-spatiales et après cette étape, aux données spatiales. (2) Approche de domination de données spatiales ; étant donné la hiérarchie de données spatiales, la généralisation est effectuée d`abord sur ces données et après sur les données non spatiales.

B.1.2 Regroupement

Le regroupement (*clustering*) est le processus permettant de regrouper de manière physique ou abstraite des objets en classes d'objets semblables. Cette approche de fouille de données nous aide à construire des groupes “représentatifs” d'un ensemble d'objets basé sur une mesure de similitude (Ex. distance euclidienne). En d'autres termes, le regroupement de données identifie des groupes (*clusters*) ou des régions densément peuplés en accord avec une certaine mesure de distance dans un ensemble de données multidimensionnelles. On peut classifier les algorithmes de regroupement en quatre groupes principaux : algorithmes de regroupement basés sur les approches *k-means* (centre de gravité du cluster) et *k-medoid* (objet représentatif du cluster), algorithmes hiérarchiques, algorithmes basés sur la localisation des objets (regroupement par densité), et dernièrement ceux qui sont basés sur des *grids*.

B.1.3 Associations spatiales

Une règle d'association spatiale est une règle qui décrit l'implication d'un ensemble d'objets vers un autre ensemble d'objets d'une base de données spatiales [29]. Un exemple d'une règle d'association spatiale peut être " si la entreprise est près de Mexico City est alors une grande entreprise". Une règle d'association spatiale est de la forme $X \rightarrow Y$, où X et Y sont des ensembles de prédicats spatiaux ou non spatiaux. Il existe plusieurs types de prédicats spatiaux qui pourraient constituer une règle d'association spatiale, par exemple, des relations topologiques comme intersection, recouvrement et des relations d'orientation spatiale comme Gauche_de, ou Ouest_de.

B.1.4 Rapprochement et agrégation

Les méthodes basées sur le rapprochement et l'agrégation cherchent à analyser les caractéristiques des groupes d'objets (*clusters*) pour former des groupes d'objets (*features*) proches entre eux. La proximité devient la mesure permettant de regrouper un ensemble de points dans un cluster en un *feature*. L'idée de trouver des relations de proximité n'est pas un problème simple comme on pourrait le croire, car il existe trois raisons pour cette affirmation. Supposons qu'on ait un cluster de points et que l'on veuille trouver les *k-features* les plus proches de lui :

- La taille et forme du cluster et les *features* peuvent être très variés.,
- On pourrait avoir une grande quantité de *features* à examiner,
- Même pour trouver une forme connue (ex. polygone) qui décrit la forme du cluster, il serait impropre de reporter les *features* où les limites soient plus proches aux

limites de celui-ci, parce que la distribution des points à l'intérieur du cluster ne peut pas être uniforme.

B.1.5 Fouille de données-images

L'extraction de patrons à partir d'images est autre approche de la fouille de données spatiales. Dans la littérature il existe divers travaux de fouille de données développés selon cette approche. Par exemple, Fayyad et al. [14] présentent un système pour l'identification et la catégorisation des volcans sur la surface de Venus à partir d'images transmises par la sonde stellaire Magellan. Dans un autre travail [15] (*Second Palomar Observatory Sky Survey*) les auteurs ont utilisé des arbres de décision pour la classification des galaxies, des étoiles et des autres objets stellaires. Stolorz et al. [44] d'une part, et Shek et al. [41] d'autre part ont effectué des fouilles de données spatio-temporelles de données géophysiques.

B.1.6 Classification de données spatiales

La classification de données spatiales a comme objectif de trouver des règles qui divisent un ensemble d'objets en un nombre de groupes dans lesquels les objets de chaque groupe appartiennent à une même classe. Divers types d'information peuvent être utilisés pour caractériser les objets spatiaux, comme par exemple, les attributs non spatiaux d'un objet, les prédicats spatiaux et les fonctions spatiales. L'idée est d'utiliser cette information pour en extraire les attributs pour l'étiquetage de classes (attributs qui divisent les données en classes) et ceux dont les valeurs sont utilisés dans un arbre de décision pour créer de nouvelles branches.

B.1.7 Détection de tendances spatiales

La détection de tendances spatiales décrit les changements réguliers d'un ou plus attributs non spatiaux d'un objet quand celui-ci se déplace d'un point de référence initiale. Un exemple de tendance spatiale serait “si on s'éloigne du centre historique de la ville de Puebla le prix des terrains décroit”. Les trajectoires de mouvement à partir d'un point *x* sont utilisées pour modéliser ce mouvement, et l'analyse de régression sur les attributs des objets peut être utilisée pour décrire les patrons de changements. Il existe deux types de tendances : globales et locales.

B.2. Motivation

Bien que les approches précitées effectuent les fouilles de données spatiales avec succès, notre perception est que celles-ci ne considèrent pas tous les éléments trouvés dans une base de données spatiales (données spatiales, données non spatiales et relations spatiales entre les objets spatiales) d'une manière exhaustive. C'est-à-dire que certaines d'entre elles d'abord réalisent la fouille de données spatiales et ensuite celles non spatiales ou vice-versa, et d'autres autorisent des combinaisons de ces éléments mais de manière restreinte. Au vu des considérations précédentes, on propose l'argument suivant : si nous sommes capables de représenter les données spatiales, non spatiales et les relations entre objets spatiales comme un unique ensemble de données, et on les fouille ainsi, on pourrait générer ou trouver des patrons de connaissances qui caractérisent notre ensemble de données contenant ces trois éléments de manière conjointe. Pour un tel objectif, on émet l'hypothèse qu'une

représentation basée sur graphes est suffisamment flexible et puissante pour représenter ces éléments de manière conjointe, facilement compréhensible et capable de créer différentes représentations du même domaine. Le domaine est décrit en utilisant des graphes, ces graphes se transformant en données d'entrée pour un outil de découverte de connaissances basé sur les graphes lequel utilise des heuristiques pour sélectionner des sous-graphes qui sont considérés comme importants (patrons).

Un graphe est défini comme une paire $G = (V,E)$. V = $\{v_1,\ldots,v_n\}$ dénote un ensemble fini d'éléments appelés sommets. E est un ensemble d'arcs e satisfaisant $E \subseteq [V]^2$. Donc, chaque arc $e \in E$ est une paire (v_i,v_j). Si (v_i,v_j) est une paire ordonnée pour n'importe quel $(v_i,v_j) \in E$, on dit que $G = (V,E)$ est un graphe orienté. Un graphe étiqueté possède des étiquettes associées à leurs sommets et aussi aux arcs.

Après avoir proposé la représentation de relations spatiales entre les objets spatiaux, dans la suivante section on détaillera les trois types de relations spatiales que l'on utilisera.

B.2.1 Relations spatiales

La position explicite des objets spatiaux définit des relations implicites de voisinage (*neighborhood*) spatial entre eux. Ainsi, l'information sur le voisinage des objets spatiaux constitue un élément de valeur qui doit être considéré comme le travail de fouille de données spatiales. Martin Ester et al. [9][11] introduisent le concept de graphes de voisinage pour représenter explicitement ces relations de voisinage implicite. Les graphes de voisinage peuvent couvrir les relations de voisinage suivantes :

- Topologiques : dérivées du modèle à neuf intersections [6][7][8], ce sont des relations que restent invariables sous des transformations linéaires, c'est-à-dire que si les deux objets simultanément tournent, se déplacent ou changent d'échelle les relations entre eux sont conservées.
- Distance : la distance entre deux objets compare à un seuil grâce à des opérateurs arithmétiques tels que <, >, =. La distance entre deux objets se définit comme la distance minimum entre eux (Ex. sélectionner tous les éléments à l'intérieur d'un rayon de 50 kilomètres d'un point *x*).
- Direction : la relation spatiale *R* de direction entre deux objets spatiaux *A* et *B* (*B R A*) est définie en utilisant un point représentatif de l'objet *A* et tous les points de l'objet but *B*. Le point représentatif de l'objet source *A* peut être le centre de l'objet ou un point sur ses limites. Ce point représentatif est utilisé comme l'origine d'un système de coordonnées virtuelles et son quadrant définit la direction.

Une fois décrite la motivation de notre travail de recherche, nous présentons à la suite le modèle général basée sur des graphes pour représenter les données spatiales.

B.3 Représentations basées sur des graphes

Comme dit auparavant, notre proposition s'appuie sur la création d'un modèle basé sur les graphes pour représenter conjointement données spatiales, non spatiales, relations spatiales entre les objets spatiaux. L'idée est d'utiliser les graphes générés comme données d'entrée pour un algorithme de fouille de données, de telle sorte que l'algorithme puisse trouver des

patrons concernant ces éléments de manière conjointe et non comme des éléments séparés. En conséquence, on propose le modèle de représentation donné en notation *UML* Figure B.1.

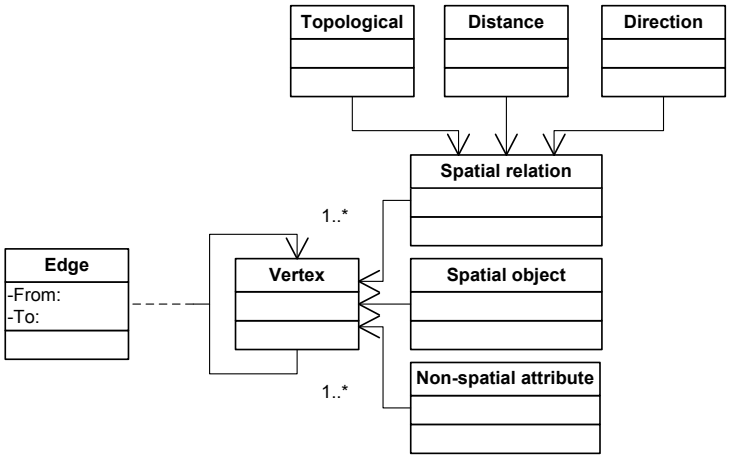

Figure B.1. Modèle basé sur des graphes pour représenter des données spatiales.

Dans ce modèle, les données spatiales (ex. objets spatiaux), données non spatiales (ex. attributs descriptifs), et relations spatiales sont représentées comme une collection d'un ou plusieurs graphes orientés étiquetés. Les sommets peuvent représenter des objets spatiaux, types de relation spatiale entre deux objets (relation binaire), ou des attributs non spatiaux décrivant les objets spatiaux. Les arcs représentent un lien existant entre deux sommets de n'importe quel type. Dépendant du type de sommets qu'un arc unit, celui-ci peut représenter le nom d'un attribut descriptif ou le nom d'une relation spatiale. Les noms des attributs peuvent se rapporter à des descriptions d'objets spatiaux et/ou à entités non spatiales. On utilise des arcs orientés pour représenter les informations directionnelles, de relations entre deux éléments (Ex. objet *x* couvre objet *y*) et pour décrire des attributs appartenant à un objet (Ex. objet *x* a attribut *z*).

Actuellement il existe cinq représentations à partir du modèle général décrit auparavant. Trois aspects définissent les caractéristiques des graphes créés dans chaque modèle: (1) Représentation des relations spatiales équivalentes (Ex. toucher, recouvrir). (2) Représentation des relations spatiales symétriques (Ex. contient-de/dans_de, Nord_de/Sud_de). (3) La manière de représenter les objets et leurs relations dans le modèle. En conséquence, les graphes créés se différencient de manière quantitative et qualitative. Dans la partie quantitative il y a des différences telles que : nombre de sommets et d'arcs employés pour représenter les données spatiales, non spatiales et les relations, création de graphes simples, l'usage des arcs orientés et non orientés pour représenter des relations spatiales et/ou des attributs descriptifs. Dans la partie qualitative on a observé, à travers des expériences, que certains modèles ont une plus grande expressivité pour représenter l'ensemble de données, conditionnant de manière directe la qualité des résultats générés dans le processus de fouille. Ensuite on présente trois des cinq modèles proposés décrivant les métriques d'évaluation créées pour caractériser à chacun d'eux.

Modèles

Dans le but de décrire les caractéristiques de chaque modèle, on va utiliser à titre d'exemple l'ensemble de données de la Figure B.2. Comme on peut l'observer, notre ensemble de données se compose de deux objets spatiaux, l'objet *A* représentant une maison et l'objet *B* représentant un lac, et les trois relations spatiales suivantes: (1) Distance, objet *A* près de objet *B* (relation équivalente). (2) Topologique, objet *A* touche objet *B* (relation équivalente). (3) Direction, objet *A* Sud de objet *B* (relation symétrique).

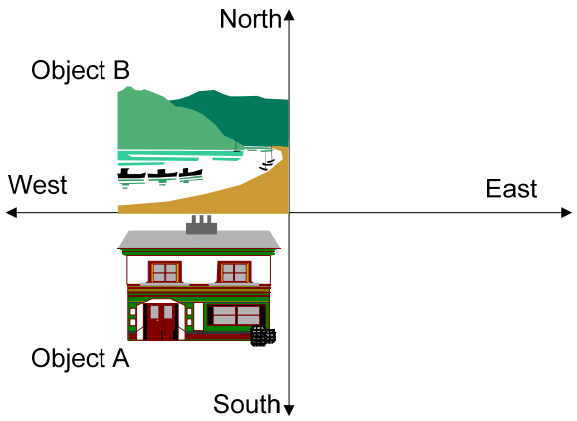

Figure B.2. Base d'exemple pour caractériser les 3 modèles proposés.

Modèle n°1 - modèle de base

La Figure B.3 montre le premier modèle créé pour représenter des données spatiales dans l'approche proposée. Les caractéristiques du modèle selon les métriques créées pour sa caractérisation sont :

- Nom. Sommets : 2 sommets, chacun représentant un objet spatial (objet *A* et objet *B*).
- Nom. Arcs : 4 arcs, 3 arcs pour représenter les relations spatiales originales existantes dans notre ensemble de données d'exemple (“près”, “touche” et “Sud_de”) et un arc pour représenter la relation “Nord_de”, créé à partir de la relation symétrique original “Sud_de”.
- Taille (sommets + arcs) : 6
- % incrément : 0%, celui-ci est le modèle base.
- Graphe simple. Non. Car c'est un graphe complexe avec 4 arcs unissant 2 sommets.
- Arc orienté. Oui, car on utilise les relations symétriques “Sud_de” et “Nord_de”. La direction des arcs se fait avec la lecture des relations entre les objets.

- Arc non orienté. Oui, car on utilise les relations équivalentes “près” et “touche”.
- Information complète. Oui, car dans le graphe est représentée la relation symétrique“Nord_de” créée à partir de la relation symétrique original “Sud_de”.
- Arc “Relation” redondant. Non, car dans le modèle on n'utilise pas les arcs “Relation”.

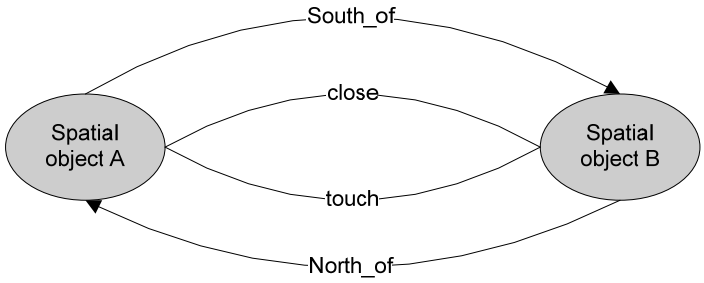

Figure B.3. Modèle n°1 - modèle base.

Modèle n°2 - réplication simple de types de relation, information complète

Dans la Figure B.4 on représente le deuxième modèle créé pour représenter les données spatiales. Les caractéristiques du modèle selon les métriques sont les suivantes :

- Nom. Sommets : 5 sommets, 2 sommets pour représenter les objets spatiales et trois sommets pour représenter les types de relations spatiales “topologique”, “distance” et “direction”. Pour chaque type de relation spéciale existant entre deux objets, on ajoute un sommet étiqueté avec le nom du type de la relation spatiale. Dans l'exemple il existe une relation “topologique”, une relation de “distance” et une relation de “direction”.
- Nom. Arcs : 6 arcs, 3 arcs pour représenter les relations originales, 2 arcs pour représenter les relations équivalentes (“près” et “touche”) créées à partir des

relations originales et un arc pour représenter la relation symétrique (“Nord_de”) créée aussi à partir des relations originales.

- Taille (sommets + arcs) : 11
- % incrément : +83.33%
- Graphe simple. Oui, car il existe un arc supplémentaire entre n'importe quel couple de sommets donnés.
- Arc orienté. Oui, car on utilise toutes les relations. La direction des arcs va des sommets représentant les objets spatiaux aux sommets représentant les types de relations spatiales.
- Arc non orienté. Non, car dans le modèle on n'utilise pas d'arcs non orientés.
- Information complète. Oui, car on représente les relations symétriques créées à partir des relations originales.
- Arc “Relation” redondant. Non, dans le modèle on n'utilise pas d'arcs “Relation”.

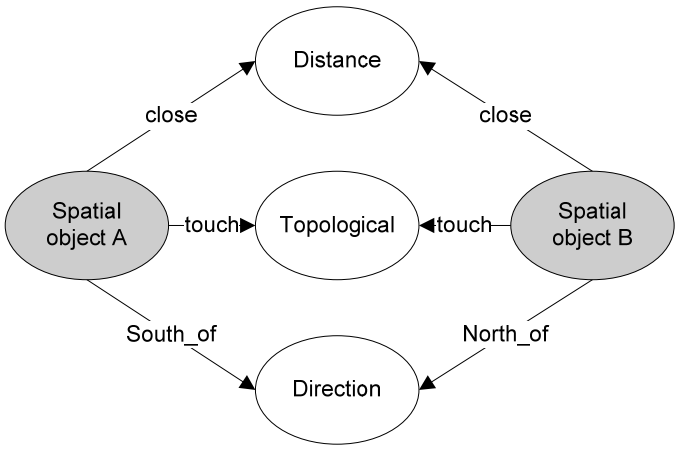

Figure B.4. Modèle n°2 - réplication simple des types de relation, information complète.

Modèle n°3 - double réplication des types de relation, information non complète

Dans la Figure B.5 on présente le troisième modèle créé pour représenter des données spatiales utilisant une approche de graphes. Les caractéristiques en accord avec les métriques sont:

- Nom. Sommets : 8 sommets, 2 sommets pour représenter les objets spatiaux et 6 sommets pour représenter les types de relations spatiales (“distance”, “topologique” et “direction”). Pour chaque type de relation spatiale entre deux objets spatiaux, on ajoute deux sommets étiquetés avec le nom du type de la relation spatiale. Par exemple, dans nos données d'essai il existe trois types de relations : 1 relation “topologique”, 1 relation “distance” et 1 relation “direction” ; ainsi on ajoute 6 sommets, 2 pour chaque type de relation spatiale.
- Nom. Arcs : 9 arcs, 6 arcs “Relation” pour unir les sommets représentant les objets spatiaux avec les sommets représentant les types de relations spatiales (de chaque sommets représentant un objet spatial naissent 3 arcs puisqu'il existe 3 types de relation), et 3 arcs pour présenter les relations spatiales originales. Ces 3 arcs sont utilisés pour unir les sommets représentant les types de relations spatiales.
- Taille (sommets + arcs) : 17
- % incrément : +183.33%
- Graphe simple. Oui, car il existe au moins un arc entre n'importe quel couple de sommets données.
- Arc orienté. Oui, car on utilise les relations symétriques et les arcs “Relation”. La direction des arcs “Relation” va des sommets représentant les objets spatiaux aux

sommets représentant les types de relations spatiales. La direction des arcs restant se fait avec la lecture des relations spatiales entre les objets.

- Arc non orienté. Oui, car on les utilise pour représenter les relations équivalentes.
- Information complète. Non, car dans le modèle ne sont pas représentées les relations symétriques qui sont créées à partir des relations spatiales originales.
- Arc "Relation" redondance Oui, car dans le modèle on utilise des arcs "Relation" pour représenter explicitement l'existence et type d'une relation spatiale entre 2 objets spatiaux. De plus, on emploie ces arcs pour éviter la création de graphes complexes.

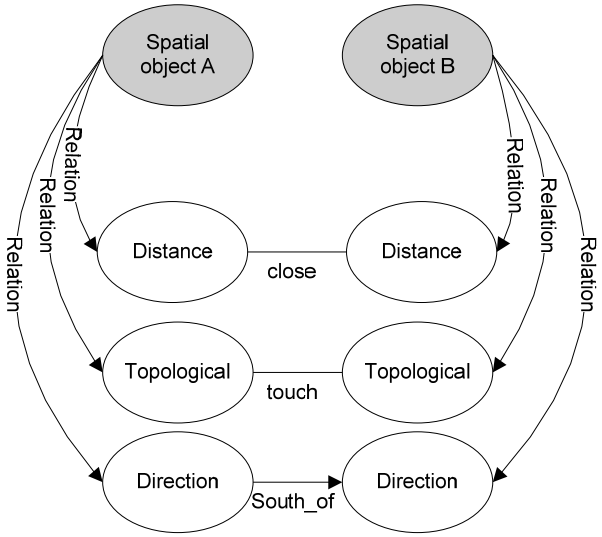

Figure B.5. Modèle n°3 - double réplication des types de relation, information non complète.

La Table B.1 présente les résultats des neuf métriques développées pour évaluer les caractéristiques de chaque modèle proposé (actuellement 5 modèles). Le modèle n°1 est appelé le modèle base.

Modèle	Nom. Sommet (1)	Nom. Arcs (2)	Dimensions (s + a) (3)	% Incrément (4)	Graphe Simple (5)	Arc Orienté (6)	Arc non Orienté (7)	Information Complète (8)	Arc "Relation" Redondant (9)
n°1	2	4	6	-	Non	Oui	Oui	Oui	Non
n°2	5	6	11	+83.33	Oui	Oui	Non	Oui	Non
n°3	8	9	17	+183.33	Oui	Oui	Oui	Non	Oui
n°4	5	6	11	+83.33	Oui	Oui	Oui	Non	Oui
n°5	8	12	20	+233.33	Oui	Oui	Non	Oui	Oui

Table B.1. Caractéristiques des modèles de représentation basés sur des graphes.

Les métriques ont été proposées sur la base de causes à effets de chacune aussi bien dans le graphe créé que dans l'algorithme de fouille. Nous apercevons quatre caractéristiques significatives liées directement à ces métriques.

1. Espace de recherche

L'espace de recherche dans un algorithme de fouille de données basé sur des graphes consiste en la liste des sous-graphes qui peuvent être dérivés du graphe initial, de telle manière que les nombres de sommets (1) et arcs (2) du graphe créé (3) définissent la taille de l'espace de recherche pour le système de découverte. Donc, l'objectif doit être de minimiser le nombre de sommets et d'arcs utilisés pour créer les graphes mais en même temps maximiser la représentativité des ceux-ci. Comme on peut le voir dans la Table B.1, le modèle utilisant le nombre minimum de sommets et d'arcs pour représenter l'ensemble de données d'exemple c'est le modèle n°1 (2 sommets et 4 arcs) alors que le modèle n°5 est le cas opposé (8 sommets et 12 arcs).

2. Temps de traitement

La taille de l'espace de recherche joue un rôle important touchant au temps de traitement utilisé pour découvrir des patrons. Si on dispose d'un grand espace de recherche, il faudra plus de temps pour évaluer tous les sous-graphes possibles. Donc, une comparaison de la métrique "pourcentage d'incrémentation" (4) entre les modèles proposés figure dans la Table B.1. Rappelons que cette métrique compare la taille d'un modèle donné par rapport au modèle n°1 (modèle base). Par exemple, le modèle n°5 augmente la taille du graphe de 233.33% par rapport au modèle n°1. C'est-à-dire que l'algorithme de fouille aura besoin d'évaluer 233.33% plus de sommets et/ou d'arcs en utilisant le modèle n°5 au lieu du modèle n°1 pour le même ensemble de données.

3. Complexité du graphe

Dans le chapitre 4 du mémoire de thèse, on décrit le système Subdue, notre outil de fouille de données basées sur des graphes, outil provenant de l'Université du Texas à Arlington. Comme dit précédemment, il existe une plus grande complexité pour l'algorithme de fouille travaillant avec des graphes complexes au lieu de graphes simples (Ex. au maximum un arc unissant n'importe quel couple de sommets donné). Par exemple, de plus, il faut tenir compte dans le traitement de comparaison des graphes, de la phase d'extension (Subdue emploie une approche d'"extension" pour découvrir des patrons), et de l'étape de compréhension du graphe. Donc, l'objectif a été de proposer des modèles basés sur les graphes qui nous permettent de créer des graphes plus simples. Comme on peut le voir dans la Table B.1, seulement le modèle n°1 ne permet pas de créer des graphes simples.

En conséquence, comme stratégie pour réduire la multiplicité du nombre d'arcs entre deux sommets, on emploie les approches suivantes:

- Ajouter un nouveau sommet étiqueté avec le nom du type de la relation (Ex. topologique, distance et direction) pour chaque relation spatiale entre les objets spatiaux. Cette approche est utilisée dans le modèle n°2.
- Ajouter un nouveau sommet (modèle n°4) ou deux nouveaux sommets (modèle n°3 et modèle n°5) étiquetés comme le nom du type de la relation spatiale (Ex. topologique, distance et direction) par chaque relation spatiale entre les objets spatiaux et fusionner ce nouveau sommet (modèle n°4) ou de nouveaux sommets (modèle n°3 et modèle n°5) avec les sommets représentant les objets spatiaux parmi des arcs étiquetés comme "Relation". L'approche employée pour fusionner les sommets diffère pour chaque modèle selon la définition de chacun d'eux. Cette nomenclature est utilisée pour représenter le fait qu'il existe une relation spatiale entre les objets spatiaux. Ces arcs sont connus comme arcs “Relation” redondants (9).

4. Représentativité des données

Les métriques arcs orientés (6), arcs non orientés (7), et information complète (8) sont utilisées pour maximiser la représentativité des données mais minimisant, aussi tant que possible, la taille du graphe et sa complexité. Les arcs orientés sont utilisés pour représenter les relations spatiales symétriques (objet *A* Nord_of objet *B*, implique, *B* Sud_of *A*), les arcs “Relation” redondantes, les attributs non spatiaux qui décrivent les objets spatiaux. Les arcs non orientés sont utilisés pour représenter les relations spatiales équivalentes (la

relation est représentée par un arc non orienté au lieu de deux arcs orientés). Finalement, l'information complète signifie que les relations spatiales symétriques entre objets spatiaux aussi sont représentées dans le modèle.

B.4 Fouille du graphe

La caractéristique de recouvrement (partage de sommets appartenant à différentes instances d'une sous-structure) accomplit un rôle important dans le système de découverte de patrons (sous-structures) dans notre outil de fouille de données basés sur les graphes à savoir le système Subdue. En conséquence, les résultats générés sont conditionnés par le fonctionnement de cette caractéristique. Pourtant, son implémentation actuelle est régulière : permettre le recouvrement entre toutes les instances d'une sous-structure (sans aucune règle) ou ne pas autoriser le recouvrement entre aucune instance appartenant à une sous-structure. C'est-à-dire, tout ou rien. Dans ce contexte, on propose une nouvelle approche appelée recouvrement limité. Un des avantages principaux de cette nouvelle approche est la capacité d'autoriser l'usager à spécifier l'ensemble de sommets où le recouvrement sera permis. Ces sommets pourraient représenter les éléments significatifs dans le contexte de travail. On donnera directement trois motivations pour proposer un nouvel algorithme, lesquelles seront expliquées dans les sous-sections suivantes:

1. Réduction de l'espace de recherche

Dans les systèmes de découverte de connaissances basés sur des graphes, l'algorithme de fouille de données utilise des graphes comme représentation de connaissance. L'espace de

recherche d'un tel algorithme consiste en tous les sous-graphes qui peuvent être dérivés du graphe initial. Le processus de découverte de sous-structures en Subdue commence par la création de sous-structures d'un seul sommet à partir du graphe d'entrée (une sous-structure par chaque étiquette de sommet qui existe au moins 2 fois dans le graphe). Dans chaque itération du processus de découverte, l'algorithme sélectionnera les meilleures sous-structures et étend les instances de ces sous-structures en ajoutant un arc voisin (ou un arc et un nouveau sommet) dans toutes les directions possibles.

Mais comme partie du processus de sélection des meilleurs substructures et donc d'extension, il existe aussi un processus de filtrage. Dans ce processus, en accord avec la valeur du paramètre de recouvrement, les instances d'une sous-structure sont évaluées : si le recouvrement est permis, les instances partageant des sommets sont maintenues ; au contraire si le recouvrement n'est pas autorisé les instances qui partagent les sommets sont écartés.

La meilleure sous-structure découverte par Subdue (en accord à les métriques d'évaluation), dans chaque itération, peut être employée pour comprimer le graphe initial qui peut donc se transformer en un nouveau graphe d'entrée pour une interaction suivante. Après plusieurs itérations, Subdue crée une description hiérarchique des données, où les substructures découvertes lors de certaines itérations peuvent être définies sur la base de sous-structures découvertes lors des itérations préalables. Ainsi, le nombre d’instances d'une substructure définit l'espace de recherche (dans chaque itération) dans le processus de découverte de sous-structures. Comme on peut l'observer à travers de l'utilisation du recouvrement limité, on obtient une réduction de l'espace de recherche puisque le nombre

d'instances candidates à être étendues conditionne les valeurs (sommets) où le recouvrement est permis.

2. Réduction du temps de traitement

La réduction du nombre d'instances candidates à être étendues apporte une réduction de l'espace de recherche, et cela contribue à une réduction du temps de traitement pour la recherche de sous-structures (patrons).

Permettre le recouvrement en Subdue, implique une évolution du temps de calcul puisqu'augmente le nombre d'instances candidates à être étendues, évaluées, comparées, et découvertes. Pourtant, avec l'implémentation du recouvrement limité, le nombre d'instances à être traitées dans ces phases décroît contribuant à une réduction du temps de traitement dans tout le calcul de découverte de sous-structures.

3. Recherche orientée de patrons avec recouvrement (partager) sélectif

Le recouvrement limité donne à l'usager la capacité de définir des ensembles d'éléments où le recouvrement sera permis et qu'à son avis il considère comme pertinents pour son contexte de travail (Ex. un objet spatial, un attribut descriptif). Ces éléments sont représentés en utilisant des sommets en accord avec le modèle proposé. En contre partie, l'algorithme écartera les éléments que l'utilisateur n'a pas considérés comme significatifs.

Par conséquent, le recouvrement limité fournit à l'utilisateur un moyen pour mettre en œuvre une recherche orientée vers des patrons avec recouvrement. C'est-à-dire, l'utilisateur délimite l'ensemble d'éléments qui auront un rôle prépondérant dans le processus de

découverte de sous-structures. De plus, cette caractéristique offre l'avantage que le processus d'évaluation de patrons est simplifié puisque l'ensemble de résultats produits est plus petit parce que ceux-ci se centrent sur les demandes de l'utilisateur.

B.5 Résultats

A partir des essais pour évaluer notre proposition de modélisation et de fouille de données spatiales en utilisant un modèle basé sur les graphes, on a développé trois cas d'utilisation exemplaires. Les deux premiers cas ont été mis en œuvre en utilisant un recensement de population du centre historique de la ville de Puebla durant l'année de 1777. Le troisième cas d'utilisation a été développé en utilisant une base de données spatiales de la région du volcan Popocatépetl.

Dans cette section nous présentons des exemples de résultats obtenus avec le cas du volcan Popocatépetl. Supposons que nous souhaitons connaître des caractéristiques communes entre les habitats, routes et rivières dans la zone qui nous aident à évaluer ou mettre en œuvre des plans d'évacuation en cas d'éventualité volcanique : par exemple, caractéristiques de routes commençant dans ou croisant un habitat, un matériel utilisé pour construire ces routes et leur état actuel (ex. pavées, non pavées), caractéristiques des routes et des rivières qui ont une certaine relation entre elles (ex. croisement), rivières près d'habitats qui en cas de haute précipitation pluviale pourraient présenter des dangers potentiels.

Les expériences ont été développées en utilisant les 5 modèles de représentation actuellement proposés. Dans cette section on présente des patrons trouvés avec le modèle n°1 entre routes et rivières, routes et habitats, et finalement rivières et habitats. L'idée d'organiser la présentation des résultats de cette manière est de montrer les divers patrons qui peuvent être découverts entre ces éléments. À la fin de la section on présente des tableaux comparant les résultats obtenus avec chacun des 5 modèles.

La Figure B.6 montre le patron le plus significatif découvert entre des routes et des rivières en utilisant le modèle n° 1. Le patron décrit une relation entre "route de catégorie non pavée qui traverse une rivière catégorie écoulement" dans la zone. Ce patron peut être considéré comme un indicateur du nombre de routes qui ont besoin d'être vérifiées en cas d'éventualité volcanique vu le type de matériel avec lequel elles sont construites, et parce qu'elles traversent des rivières (la lecture peut être faite en sens inverse) qui en cas de hautes concentrations pluviales peuvent déborder et les mettre hors d'usage. Subdue a trouvé avec non recouvrement 46 instances du patron dans la seconde itération ; par l'intermédiaire du recouvrement standard il a trouvé 85 instances dans la première itération ; et à travers le recouvrement limité il a aussi trouvé 85 instances dans la seconde itération. Comme nous pouvons observer les recouvrements standard et recouvrement limité ont trouvé le même nombre d'instances, mais le recouvrement limité a eu besoin de deux interactions pour trouver le même patron. Toutefois, ceci ne veut pas dire que le recouvrement standard est meilleur que le recouvrement limité (en ce qui concerne temps de traitement) parce qu'en analysant le temps global de traitement requis pour le recouvrement limité pour achever la phase de découverte de sous-structures nous remarquons qu'il est plus petit que celui requis par le recouvrement standard.

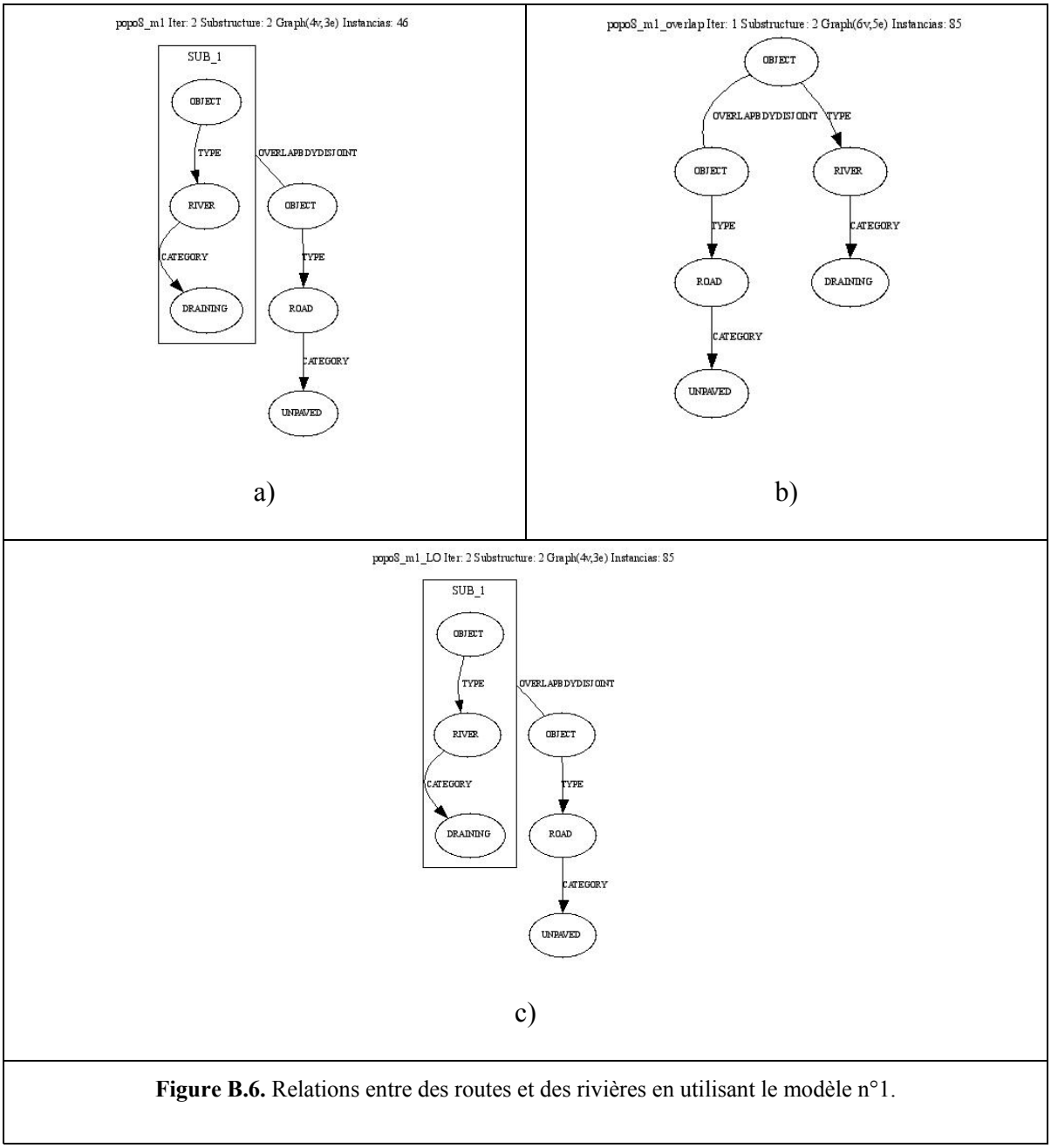

Figure B.6. Relations entre des routes et des rivières en utilisant le modèle n°1.

Le patron le plus significatif, utilisant le modèle n°1, trouvé entre des routes et des habitats est présenté dans la Figure B.7. Il décrit une relation entre "route catégorie non pavée en touchant un habitat catégorie construction". "habitat catégorie construction" représente dans la couche de données spatiales "habitat" de la base de données du volcan, habitats avec grosse population, bâtiments et une grande quantité de constructions utilisées pour offrir

des services aux habitants. Si nous faisons l'hypothèse que les gens pourraient avoir besoin d'être évacués en cas d'éruption et que les routes utilisées pour ce but sont non pavées, alors, cette situation pourrait se transformer en un problème (Ex. embouteillage). Dans cette expérience Subdue a trouvé par l'intermédiaire du non recouvrement 6 instances du patron dans la neuvième itération ; par le biais du recouvrement standard 9 instances dans la quatrième itération ; et en utilisant recouvrement limité 8 instances dans la dixième itération.

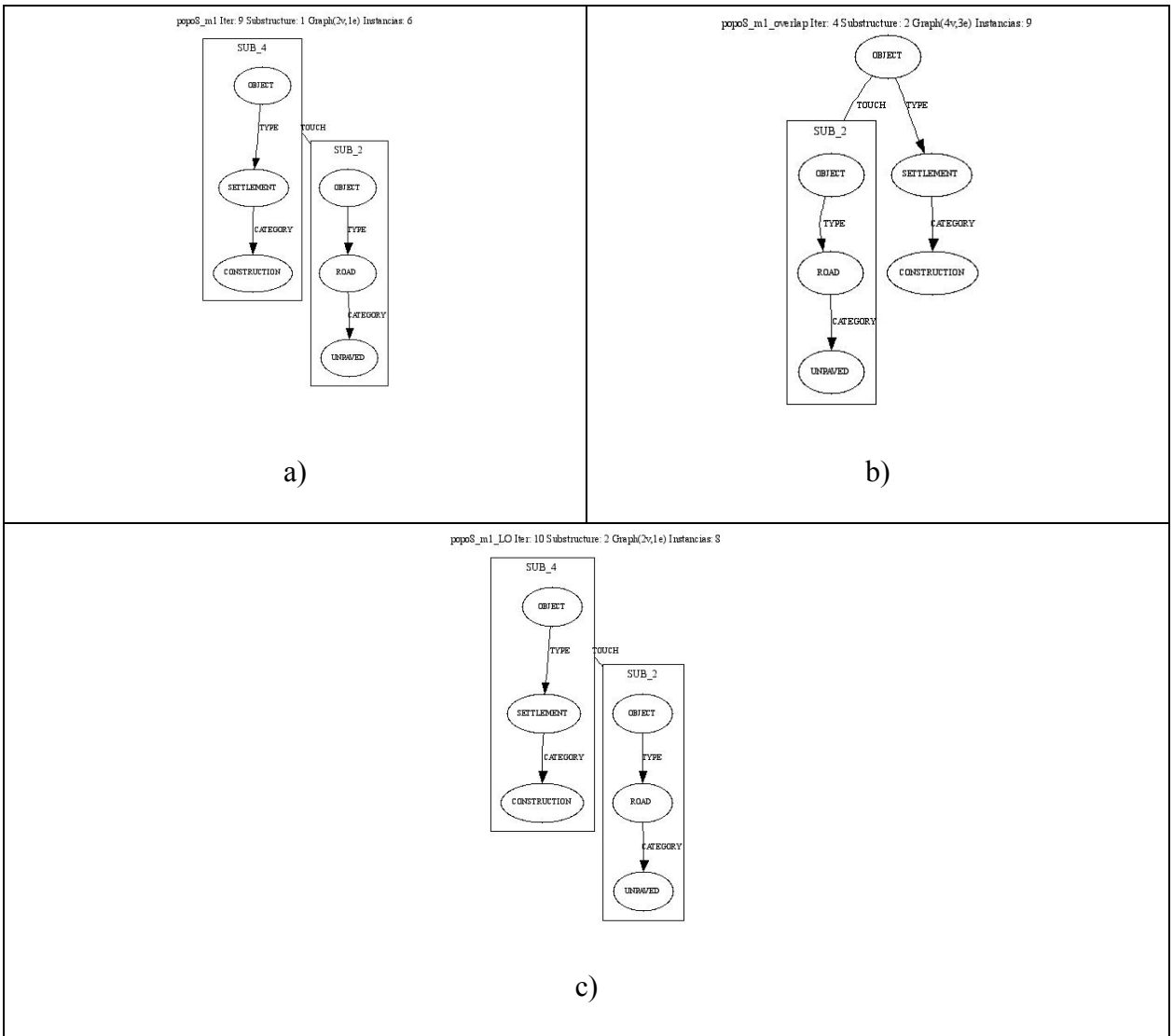

Figure B.7. Relations entre des routes et des villes en utilisant le modèle n°1.

La Figure B.8 montre le patron le plus significatif trouvé entre des rivières et des populations en utilisant le modèle n° 1. Le patron décrit une relation entre "rivière catégorie écoulement qui traverse un habitat catégorie "îlot" ou "bâtiment" dans la zone. "habitat catégorie îlot" représente dans la couche de données spatiales "habitat" de la base de données du volcan, des villages avec peu de population, en fait avec beaucoup de secteurs dépeuplés, bâtiments et constructions précaires. Le patron peut être utilisé pour identifier des zones potentielles d'inondation, habitées par des personnes isolées habitant aux alentours de rivières. À travers le non recouvrement, Subdue a trouvé 5 instances du patron dans la dixième seconde itération ; en utilisant le recouvrement standard a trouvé 5 instance dans la huitième itération ; et par l'intermédiaire du recouvrement limité il a aussi trouvé 5 instances en huitième itération. Subdue a trouvé, toutefois, le même patron dans les trois cas mais en utilisant le recouvrement standard et le recouvrement limité.

Le Tableau B.2 présente une comparaison, par modèle, entre le nombre d'instances découvertes/itérations nécessaires pour les découvrir et les trois mises en œuvre du recouvrement. Par exemple, en utilisant le modèle n°1, Subdue a trouvé 46 instances (dans la seconde itération) d'un patron "complet" (notre définition pour décrire un patron "complet" est que celui-ci contient au moins deux objets spatiaux et la relation spatiale entre eux) en contenant les objets spatiales route-rivière par l'intermédiaire du non recouvrement. Une valeur plus élevée signifie un modèle qui permet de découvrir plus d'instances d'une sous-structure (patrons). Rappelons que Subdue décrit comme le meilleur patron (par itération) la sous-structure avec le nombre plus élevé d'instances découvertes de

cette sous-structure. Cette comparaison est effectuée par chaque structure "objet-objet" (Ex. route-rivière).

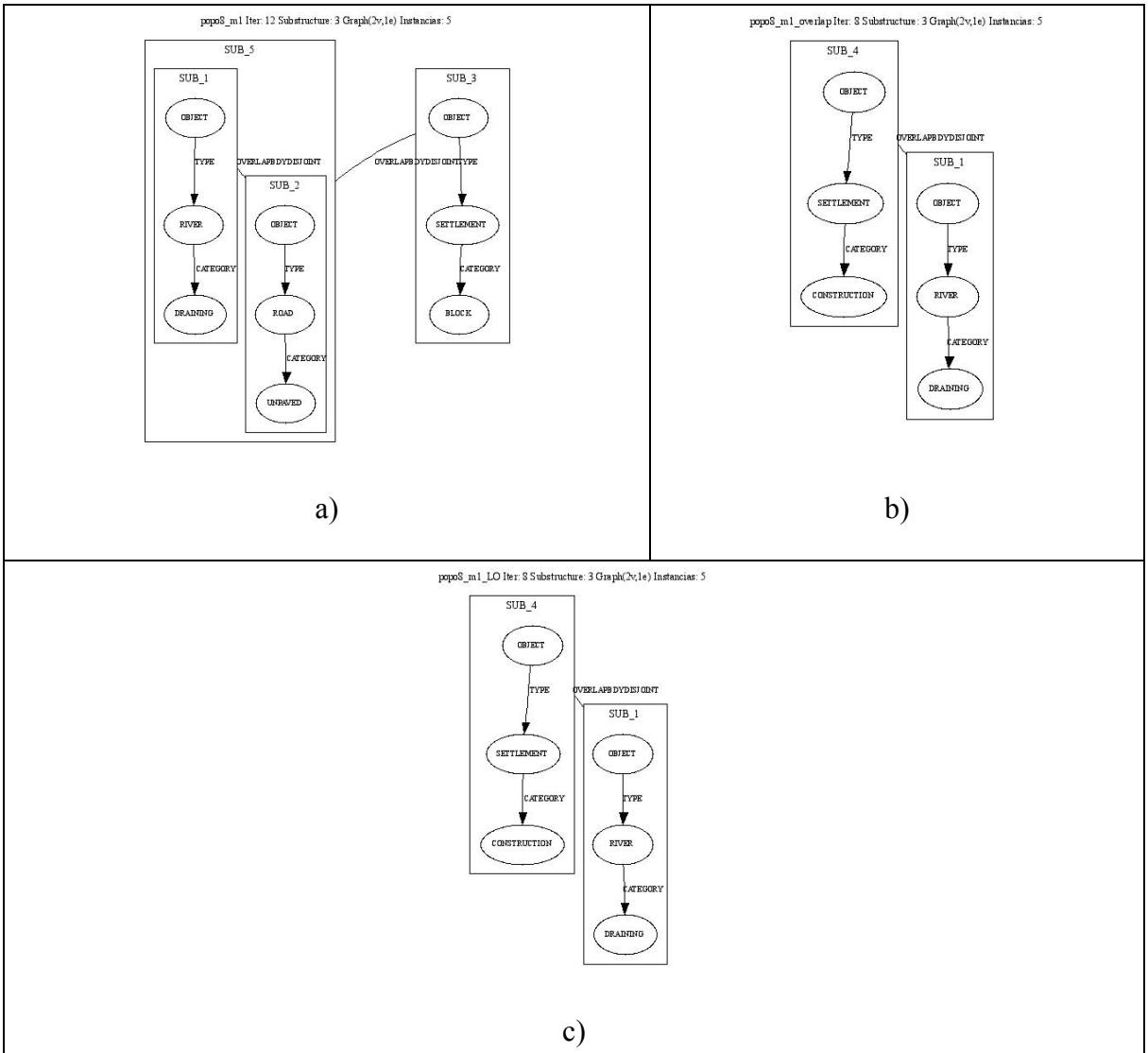

Figure B.8. Relations entre des rivières et des villes en utilisant le modèle n°1.

Annotation: NO (non recouvrement), SO (recouvrement standard), LO (recouvrement limité).

	Modèle n°1			**Modèle n°2**			**Modèle n°3**			**Modèle n°4**			**Modèle n°5**		
	NO	**SO**	**LO**	**NO**	**SO**	**LO**	**NO**	**SO**	**LO**	**NO**	**SO**	**LO**	**NO**	**SO**	**LO**
Route-Rivière															
Instances	46	85	85	41	85	64	39	85	34	39	85	60	45	85	45
Itération	2	1	2	2	3	2	2	2	9	2	1	2	5	1	5
Route-Ville															
Instances	6	9	8	5	8	7	4	8	5	6	8	8	6	0	7
Itération	9	4	10	14	6	10	15	6	13	12	10	7	7	0	7
Rivière-Ville															
Instances	5	5	5	5	10	5	5	19	5	5	10	5	5	5	5
Itération	12	8	8	16	7	14	12	4	10	6	6	11	13	6	10

Table B.2. Instances/itérations par chaque modèle basé sur des graphes : cas d'utilisation Popocatépetl.

Le Tableau B.3 présente une comparaison de maximum/minimum instances découvertes par chaque mise en œuvre du recouvrement. Un modèle avec la valeur plus élevée est meilleur parce qu'il permet de découvrir davantage d'instances d'une sous-structure. La comparaison est présentée par chaque structure "objet-objet". Par exemple dans la structure route-rivière le modèle n°1 a trouvé 46 instances découvertes dans la seconde itération (la valeur plus élevée).

	Maximum	Minimum
Route-Rivière		
Non recouvrement	modèle n°1 (deuxième itération)	modèles n°3 et n°4 (deuxième itér.)
Recouvrement standard	modèles n°1, n°4 et n°5 (première itér.)	modèle n°2 (troisième itération)
Recouvrement limité	modèle n°1 (deuxième itération)	modèle n°3 (neuvième itération)
Route-Ville		
Non recouvrement	modèle n°5 (septième itération)	modèle n°3 (quinzième itération)
Recouvrement standard	modèle n°1 (quatrième itération)	modèle n°5 (modèle non complet)
Recouvrement limité	modèle n°4 (septième itération)	modèle n°3 (treizième itération)
Rivière-Ville		
Non recouvrement	modèle n°4 (sixième itération)	modèle n°2 (seizième itération).
Recouvrement standard	modèle n°3 (quatrième itération)	modèle n°1 (huitième itération).
Recouvrement limité	modèle n°1 (huitième itération)	modèle n°2 (quatorzième itération)

Table B.3. Max/Min d'instances découvertes par "objet-objet"/caractéristique recouvrement.

Le Tableau B.4 présente une comparaison entre la moyenne d'instances découvertes par modèle. Une valeur plus élevée signifie un modèle permettant de découvrir plus d'instances d'une sous-structure. Chaque valeur représente la moyenne de sous-structures découvertes en utilisant le non recouvrement, le recouvrement standard et le recouvrement limité. La comparaison est donnée par chaque structure "objet-objet".

	Modèle n°1	Modèle n°2	Modèle n°3	Modèle n°4	Modèle n°5
Route-Rivière	72.0	63.3	52.7	61.3	58.3
Route-Ville	7.7	6.7	5.7	7.3	4.3
Rivière-Ville	5.0	6.7	9.7	6.7	5.0

Table B.4. Moyenne d'instances découvertes par modèle/"objet-objet".

Le Tableau B.5 présente une comparaison entre la moyenne d'instances découvertes par modèle. Une valeur élevée signifie un modèle permettant de découvrir plus d'instances d'une sous-structure. La comparaison est donnée par chaque mise en œuvre du recouvrement.

Modèle n°1			Modèle n°2			Modèle n°3			Modèle n°4			Modèle n°5		
NO	SO	LO	NO	SO	LO	NO	SO	LO	NO	SO	LO	NO	SO	LO
19.0	33.0	32.7	17.0	34.3	25.3	16.0	37.3	14.7	16.7	34.3	24.3	18.7	30.0	19.0

Table B.5. Moyenne d'instances découvertes par modèle/caractéristique recouvrement.

Le Tableau B.6 présente une fin comparative entre la moyenne d'instances découvertes par modèle. Nous pouvons voir dans le tableau que le modèle n°1 donne la valeur plus élevée d'instances découvertes (en accord avec nos paramètres des instances complètes) dans ce cas d'utilisation exemplaire. Les modèles suivants sont le modèle n°2 et le modèle n°4 respectivement.

Modèle n°1	Modèle n°2	Modèle n°3	Modèle n°4	Modèle n°5
28.2	25.6	22.7	25.1	22.6

Table B.6. Moyenne d'instances découvertes par modèle.

B.6 Conclusions

L'interaction constante entre les êtres humains et leur habitat naturel, la planète terre, produit jour après jour, de nouvelles demandes associées à la manipulation et l'exploitation

de données spatiales. Par exemple, l'analyse urbaine, la prévention les risques naturels, l'exploration de l'espace stellaire, la pollution des océans, et le déboisement des sols, pour nommer certains d'entre eux. La fouille de données spatiales intègre l'intégration de méthodes et techniques provenant de divers domaines scientifiques lesquels nous aident, au moyen d'algorithmes d'analyse et de découverte, à produire une énumération particulière de patrons sur les données spatiales.

Notre argumentation dans ce mémoire de thèse se base sur l'idée que la fouille de données spatiales ne considère pas tous les éléments trouvés dans une base de données spatiales (données spatiales, données non spatiales et relations spatiales entre les objets spatiales) d'une manière exhaustive. Par conséquent, on a proposé d'employer une analyse basée sur les graphes pour représenter ces éléments comme un unique ensemble de données, les fouiller comme un tout, de façon à pouvoir découvrir des patrons contenant les deux types données et relations spatiales (patrons les plus descriptifs).

Dans notre modèle les relations spatiales entre les objets spatiaux sont incluses parce qu'une caractéristique significative des données spatiales est l'influence que les voisins d'un objet peuvent avoir avec l'objet lui-même. Dans notre modèle nous incluons trois types de relations spatiales. A partir du modèle général on a proposé cinq modèles opérationnels. Trois aspects définissent les caractéristiques d'un graphe créé avec ces modèles : (1) Représentation des relations spatiales équivalentes. (2) Représentation de relations spatiales symétriques. (3) La manière de représenter les objets et leurs relations. Comme partie intégrante de notre méthodologie pour la fouille de données spatiales en utilisant une analyse basée sur les graphes, nous utilisons le système Subdue comme outil de fouille.

Nous avons proposé un nouvel algorithme appelé recouvrement limité lequel donne à l'utilisateur la capacité de spécifier l'ensemble de sommets sur lesquels le recouvrement est permis. Nous donnons trois motivations pour proposer cette nouvelle analyse: (1) Réduction de l'espace de recherche. (2) Réduction du temps de processus. (3) Recherche orientée de patrons avec recouvrement (partager) sélectif.

Pour démontrer la viabilité, capacité de fouille et de découverte de patrons en utilisant l'analyse proposée, on a développé un prototype pour la mise en œuvre de notre modèle en créant les ensembles de données basées sur les graphes, pour fouiller ces graphes (par l'entremise du système Subdue) et pour visualiser les patrons découverts. Les résultats produits des cas d'utilisation développés nous donnent un vaste panorama de ce que nous pourrions obtenir en utilisant cette analyse. Il est important de généraliser le fait que nous pouvons utiliser cette méthodologie de modélisation et de représentation dans tout domaine qui peut être représenté pour un graphe.

Les perspectives d'amélioration de notre travail (modèle basé en graphes, algorithme de fouille de données et système prototype) incluent les points suivants :

- **Visualisation de connaissances découvertes.** Par exemple, visualisation de résultats sur les couches spatiales, à travers l'utilisation d'icones, et la navigation dans la hiérarchie de patrons découverts.
- **Amélioration des algorithmes employés pour créer les ensembles de données basées sur des graphes en accord avec les modèles proposés.** La validation des

relations spatiales entre objets spatiaux est une phase qui dans la majorité des cas requiert une grande quantité de ressources en calcul.

- **Fouille de graphes.** On a employé le système Subdue comme outil de fouille de données. De plus, on a proposé un nouvel algorithme appelé recouvrement limité. L'isomorphisme de graphes est un problème NP-complet et par conséquent, les algorithmes devront être capables de réduire les temps de traitement pour la recherche de patrons.
- **Manipulation et représentation de relations entre des données non spatiales.** Les relations implicites et explicites entre les attributs en décrivant les objets spatiaux peuvent être inclus dans le modèle afin d'améliorer la représentation des données.

B.7 Contribution

La contribution à la découverte de connaissances dans le domaine des données spatiales, décrite dans ce mémoire, est le fruit d'une nouvelle façon de modéliser et de fouiller les données spatiales en utilisant une représentation basée sur des graphes. Cette approche inclut les aspects suivants :

- Nous avons proposé une nouvelle représentation de données basée sur les graphes pour traiter les données spatiales. On a atteint deux objectifs pour créer un modèle de données avec ces caractéristiques. Le premier d'entre eux est de créer un seul ensemble de données, basé sur des graphes, en représentant ces éléments en rapport les uns avec les autres. Le deuxième est d'employer cet ensemble de données pour

alimenter un système de fouille de données basé sur des graphes, de telle sorte que puissions-nous découvrir des patrons contenant des données spatiales, non spatiales et des relations spatiales lesquelles nous aident à décrire et comprendre les données, tout ceci basé la prémisse qu'il s'agit d'éléments en rapport dans le monde réel.

- Nous avons proposé un nouvel algorithme pour découvrir des sous-structures (patrons) en utilisant une analyse de recouvrement limité dans le système Subdue. Nous donnons directement trois motivations pour proposer la mise en œuvre du nouvel algorithme : réduction de l'espace recherche, réduction du temps de processus et recherche orientée de patrons avec recouvrement sélectif (specialized overlapping pattern oriented search).
- On a conçu et on a mis en œuvre un prototype pour le modèle proposé. Le prototype offre une interface d'utilisateur convivial pour la manipulation des couches spatiales avec lesquelles on travaillera, pour la création de graphes spatiaux et non spatiaux, pour la fouille de ces graphes (à travers du système Subdue) et pour le déploiement des résultats produits.