
Le Data Mining Spatial et les bases de données spatiales

Karine Zeitouni — Laurent Yeh

*Laboratoire PRISM - Université de Versailles - Saint-Quentin,
45, Avenue des Etats-Unis, F-78035 Versailles cedex
Karine.Zeitouni@prism.uvsq.fr, Tsin-Shu.Yeh@prism.uvsq.fr*

RÉSUMÉ. La combinaison du Data Mining et de bases de données spatiales (BDS) offre de nouvelles perspectives pour l'analyse spatiale des données géographiques. Cette combinaison amène au domaine du Data Mining Spatial (DMS). Cet article décrit dans une première partie les techniques spécifiques de gestion de bases de données spatiales dans un Système d'Informations Géographiques (SIG). La seconde partie introduit le concept de Data Mining Spatial en soulignant sa spécificité par rapport au Data Mining sur des bases relationnelles. Elle souligne l'existence de deux approches au DMS, l'une issue du domaine d'apprentissage sur des bases de données et l'autre des statistiques orientées analyse spatiale. Les méthodes développées dans l'une et l'autre de ces deux approches sont décrites et classées par catégorie selon leur rôle ou la forme de connaissance extraite. L'étude comparative de ces approches montre leurs similarités, leurs différences et leur complémentarité. Elle permet de conclure, d'une part, que la combinaison des deux approches de DMS serait profitable dans le processus d'analyse et, d'autre part, que les relations spatiales jouent un rôle central dans le processus de Data Mining et par conséquent que les BDS qui permettent de les déterminer sont d'autant plus importantes.

ABSTRACT. The combination of Data Mining and Spatial Database (SDB) offer new prospects for spatial analysis in geographical applications. This combination brings to the field of Spatial Data Mining (SDM). This paper describes, in a first part, the specific techniques for the management of spatial databases in a Geographical Information System (GIS). The second part introduces the concept of Spatial Data Mining by underlining its specificity with regard to Data Mining on relational databases. We point out that there exists two approaches for the SDM. The first comes from spatial database learning, while the second is based on spatial statistics field. Methods developed in both approaches are described and classified by category according to their role or the type of extracted knowledge. The comparative study of these two approaches shows their similarities, their differences and their complementary. We conclude first, that the combination of the two SDM approaches would be valuable in the analysis process; second that spatial relationships play a central role and accordingly, SDB that allow their computation are all the more interesting.

MOT-CLÉS : Data Mining spatial, analyse spatiale, relations spatiales, statistiques spatiales, bases de données spatio-temporelles, systèmes d'informations géographiques.

KEYWORDS: Spatial Data Mining, Spatial Analysis, Spatial Relationships, Spatial Statistics, Spatio-Temporal Databases, Geographic Information Systems.

1. Introduction

Avec le développement de la cartographie numérique, le volume de données dans les bases de données spatiales ne cesse d'augmenter. Ces données sont de plus en plus utilisées dans des applications décisionnelles, surtout depuis le développement d'outils de géocodage permettant la localisation par l'adresse. C'est le cas en géomarketing, dans l'analyse de la criminalité ou dans l'analyse de risques d'accidents ou d'épidémies. Seulement, la nature et le volume de données de base dépassent les capacités humaines d'analyse. D'où l'intérêt d'appliquer des techniques d'extraction automatique de connaissances telles que le Data Mining aux bases de données géographiques.

Le Data Mining spatial (DMS) est né du besoin d'exploitation dans un but décisionnel de données à caractère spatial produites, importées ou accumulées, susceptibles de délivrer des informations ou des connaissances par le moyen d'outils exploratoires (de fouille de données). Il constitue un domaine à part car il considère les interactions des objets dans l'espace. Ce domaine intègre des techniques provenant à la fois des bases de données spatiales et des SIG, du Data Mining et des statistiques spatiales.

L'objectif de cet article est de donner un aperçu de chacune de ces techniques dans la perspective du Data Mining spatial en montrant leur rôle dans l'analyse spatiale des données. Cette étude de l'état de l'art regroupe deux approches d'analyse spatiales, l'approche statistique et l'approche issue de travaux en bases de données spatiales. Elle propose une comparaison de ces deux approches.

En premier lieu, cet article décrit la problématique de l'analyse spatiale en s'appuyant sur un exemple d'application dans le domaine de l'analyse du risque routier. La section 3 est une introduction aux bases de données spatiales, montrant leurs capacités, leurs limites d'analyse, de même que leur utilité dans le processus de DMS. Après une brève introduction au Data Mining, l'état de l'art sur le Data Mining spatial sera exposé dans la quatrième section, en explicitant ses différentes méthodes. La dernière section fait une synthèse de ces travaux et une comparaison des approches collectées dans cet état de l'art.

2. Problématique de l'analyse spatiale

Cette partie décrit la spécificité et les objectifs de l'analyse spatiale. Pour cela, elle commence par exposer un exemple d'application traité dans le cadre du projet PSIG¹ dans lequel nous sommes impliqués, ensuite elle souligne les insuffisances des modes d'analyse traditionnels.

1. Programme Système d'Information Géographique soutenu par le CNRS (SHS et SPI) et l'IGN.

2.1. Exemple de l'analyse du risque d'accidents routiers

De plus en plus de villes et de départements disposent de données numérisées sur les accidents routiers, sur le réseau routier et parfois sur le flux de véhicules et même la mobilité. Ces données sont recueillies par les administrations ou par les services de police et de gendarmerie. De plus, cet effort s'amplifie au fur et à mesure de la mise en place du fichier national BAAC². Par ailleurs, d'autres bases de données fournissent des informations complémentaires sur l'environnement géographique comme le découpage administratif en communes, quartiers ou îlots, le bâti, la population, les équipements, etc. Ces données sont localisées et certaines, comme les accidents, sont datées.

Ces données, constituant une base de données géographique, renferment une mine d'informations utiles pour l'analyse du risque d'accidents. Cela a motivé tout naturellement l'application de techniques d'extraction de connaissances sur ces données spatiales — Data Mining spatial — dans deux programmes de recherche PSIG successifs³.

L'objectif de cette application se résume en deux fonctions. La première est d'identifier des zones à risque en analysant la répartition spatiale des accidents. La seconde est d'expliquer ce risque en recherchant des correspondances avec des propriétés de l'environnement géographique [ZEI 98]. Ces fonctions sont à caractère spatial. Or, l'approche traditionnelle pour évaluer ce risque utilise l'analyse des données attributaires (alphanumériques). Elle ne peut donc offrir de telles fonctions.

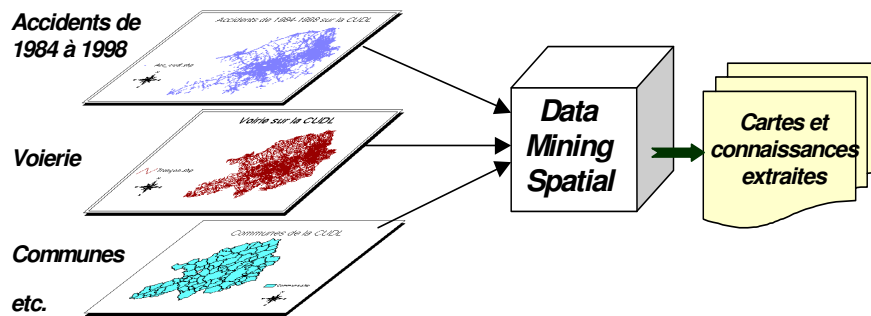


Figure 1. Data Mining spatial en analyse de risque d'accidents de la route

2. BAAC signifie « Bulletin statistique d'Analyse sur les Accidents Corporels ».

3. Projet PSIG 1998 « Extraction de connaissances des bases de données spatiales en accidentologie routière » et projet PSIG 1999 « Data Mining spatial en accidentologie routière : développement de méthodes explicatives ».

Le projet est basé sur les données de la communauté urbaine de Lille, lesquelles décrivent des accidents sur une période assez longue et donnent d'importantes informations urbaines et des informations sur la voirie (cf. figure 1).

2.2. Limites des analyses traditionnelles

Aujourd'hui, l'analyse de données en géographie s'appuie essentiellement sur l'analyse de données multidimensionnelles sans tenir compte des données spatiales [SAN 89]. Cette analyse peut être effectuée par diverses méthodes, des plus simples comme les statistiques élémentaires (moyenne, variance, histogramme...), à l'analyse multidimensionnelle — plus exploratoire — basée sur l'analyse factorielle, en passant par l'analyse bi-variée (corrélations, régressions). Ces méthodes s'appliquent toutes à des données quantitatives ou qualitatives et non à des données spatiales. De plus, elles supposent les observations indépendantes. Par conséquent, l'autocorrélation spatiale qui est une propriété majeure des données spatiales fait défaut.

Ce n'est que récemment que certains SIG ou des outils de statistiques (notamment Splus Spatial Stat [MAT 98]) commencent à intégrer des techniques de géostatistique et de statistiques spatiales. En outre, l'interaction des analyses statistiques avec la carte permet de tirer profit du pouvoir d'expression de la visualisation dans les SIG. Ces possibilités sont offertes dans les outils récents comme Spatial Analyst dans ARCVIEW.

Par ailleurs, l'interrogation des bases de données spatiales offre une forme d'analyse spatiale en formulant des requêtes (voir section 3.2). Ces requêtes peuvent inclure les relations spatiales entre objets mais elles ne sont pas exploratoires car elles sont guidées par l'utilisateur. Par exemple, l'utilisateur peut interroger la base sur les communes de plus de 10 000 habitants où le taux d'accidents par nombre d'habitants est supérieur à la moyenne. De même, l'analyse globale est très limitée par manque de calculs avancés ou spécifiques à l'analyse spatiale (indice de Geary, K-fonctions). Ces fonctions seront explicitées dans la section 4.3. Enfin, les requêtes ne découvrent ni modèles, ni règles ou relations spatiales. Néanmoins, cette approche peut être intéressante pour filtrer les données, extraire et structurer les données sur lesquelles l'analyse devra se focaliser. A ce titre, elles font partie du processus global d'extraction de connaissances.

2.3. Besoins

Les besoins de l'analyse de données spatiales aujourd'hui peuvent être classés en trois volets. Le premier est d'explorer les bases de données spatiales afin d'y découvrir des connaissances cachées. De nombreuses sources de données existent. Initialement conçues pour un autre usage, elles renferment potentiellement des informations utiles que l'analyse exploratoire permet d'extraire.

Le second est de gérer de gros volumes de données tout en garantissant des temps de réponse acceptables. En effet, l'avènement d'outils de saisie et de communication facilite l'acquisition de diverses sources de données, rendant les bases de plus en plus volumineuses. Pour que ce volume n'affecte pas les performances de traitements, des méthodes d'optimisation sont nécessaires.

Le troisième volet, qui fait la spécificité de ce domaine, est de prendre en compte les interactions dans l'espace. En effet, les données géographiques forment un continuum spatiotemporel et les propriétés d'un endroit sont souvent liées aux propriétés de l'entourage, voire expliquées par celles-ci. Ces relations dans l'espace sont caractéristiques des données géographiques (1^{ère} loi en géographie [TOB 79]). On distingue deux types de relations. D'un côté, celles liant les valeurs dans une même classe d'objets ou monothème. Par exemple, la matrice de voisinage sur les communes, utilisée dans le test d'autocorrélation spatiale. De l'autre côté, celles traduisant une relation avec les propriétés des autres couches thématiques (multithème), par exemple, le lien d'une caractéristique de l'accident et d'un type de route ou de quartier.

3. Introduction aux bases de données spatiales

Dans le domaine des Systèmes d'Informations Géographiques (SIG), les progrès de ces dernières années ont permis de dégager un certain nombre de notions couramment admises. Cette section introduit ces notions caractérisant les bases de données spatiales. Nous présenterons les fonctionnalités spécifiques d'un SIG pour ensuite introduire les aspects de modélisation des données spatiales.

3.1. Les Systèmes d'Informations Géographiques

Les SIG sont issus de technologies diverses dont l'infographie et les Systèmes de Gestions de Bases de Données (SGBD).

3.1.1. Caractérisation et fonctions d'un SIG

Un SIG se caractérise par cinq fonctionnalités connues sous le nom des 5A, à savoir : Acquisition, Assemblage, Archivage, Analyse et Affichage des données géographiques. Un SIG est avant tout un Système d'Information (SI) au sens large capable d'assurer ces 5A [LAU 94]. Ces fonctionnalités illustrées sur la figure 2 correspondent aux capacités suivantes :

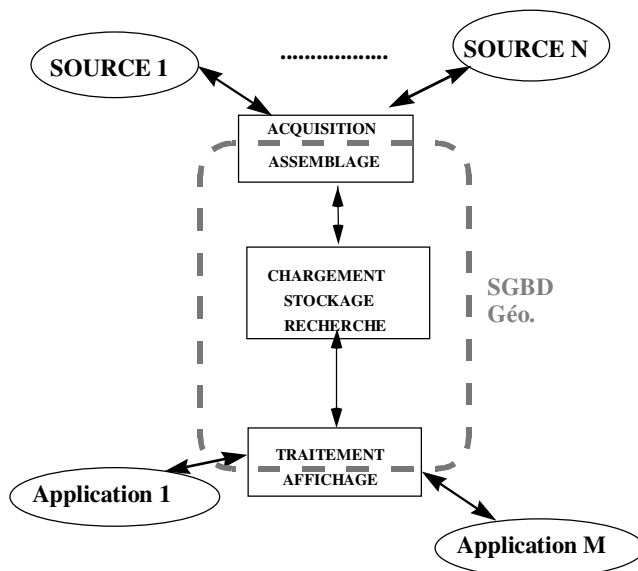


Figure 2. Les 5A et les étapes de traitements dans un SIG

– *Acquérir des données géographiques.* Cela va du simple moyen de saisir quelques données spatiales jusqu’à pouvoir générer un lot de données complet dans un processus semi-automatique.

Pour le premier, l’acquisition de la donnée s’entend souvent de manière ponctuelle. Ce sont les données spécifiques d’une application qui sont produites par l’organisme exploitant le SIG. Par exemple, l’on peut, sur un fond de carte, vouloir ajouter des données ponctuelles représentant des accidents. Une méthode de saisie peut utiliser, par exemple, une table à digitaliser.

Quant à la constitution d’un lot de données complet, c’est une tâche ardue, longue et coûteuse économiquement. C’est pourquoi peu d’entreprises ou d’institutions sont en mesure de générer leurs propres données géographiques. Cette tâche reste du ressort de producteurs comme l’IGN pour la France. Elle utilise diverses techniques telles que la photogrammétrie et le traitement d’images sur des données satellitaires. Les exemples de lots connus sont : BD-Carto de l’IGN qui décrit la France à l’échelle 1/50 000 et occupe 10 Go et la BD-Topo de l’IGN qui est à l’échelle 1/25 000 qui occupe plus de 100 Go.

– *Assembler les données géographiques.* Ceci consiste à importer des données de sources différentes et à les fusionner afin de constituer la base de données géographique adaptée aux besoins du SIG de l’entreprise.

Comme un SIG exploite souvent des données produites à l'extérieur de l'entreprise qui le développe, cette étape est plus importante que dans un système d'information classique. De plus, comme ces données sont disponibles dans l'un des multiples formats ou standards d'échange de données géographiques, elle est bien plus complexe. Parmi ces standards, citons les normes DIGEST-VPF (OTAN), EDIGéo (AFNOR) ou les formats liés aux produits comme DXF (AutoCad), MIF/MID (MapInfo), ou SHP (ArcView). Par conséquent, le chargement de ces données passe le plus souvent par des conversions de formats.

Hormis le chargement, l'assemblage de données multisource nécessite leur fusion. Si l'on intègre dans cette opération la détection des redondances et la vérification de la cohérence des données, non seulement on retrouve les problèmes d'intégration de schémas connus en bases de données traditionnelles et notamment relationnelle, mais en plus, des problèmes de précision géométrique et d'incompatibilité d'échelles s'y ajoutent. Par exemple, des écarts géométriques, dus à l'imprécision, peuvent apparaître entre deux lignes correspondant à un même objet dans deux sources différentes. Cela n'est pas aisément détecté au moment de la fusion.

– *Archiver ces données.* Ceci correspond aux fonctionnalités allant de simples systèmes de gestion de fichiers aux Systèmes de Gestion de Bases de Données (SGBD).

Les premiers ont l'avantage d'une mise en œuvre facile mais sont faibles sur les aspects performance ou fiabilité pour le stockage et le filtrage des données. Par contre, ces problèmes sont bien maîtrisés dans le cadre des SGBD.

– *Analyser ces données géographiques.* Pour une telle analyse, des outils statistiques peuvent être employés. Une catégorie d'analyse ou de préparation à l'analyse peut correspondre à l'extraction sélective suivant des critères spatiaux de données géographiques d'une base. En d'autres termes, des cartes à la demande suivant une sémantique spatiale. Par exemple, la recherche des sections de routes les plus dangereuses au sens où le nombre de blessés est plus grand que la moyenne.

– *Afficher des données géographiques.* Cet affichage n'est pas un simple problème de dessin vectoriel comme le font de nombreuses bibliothèques graphiques. En effet, l'affichage est riche de symboliques pour représenter des cartes avec des légendes. Un exemple simple concerne le tronçon de route qui, dans une base de données, est géométriquement représentée par des polylignes. Mais pour l'utilisateur, en fonction du type de route (autoroute, chemin départemental), l'affichage apparaît sous différents aspects. Un tronçon d'autoroute apparaît en gras et doublé, un tronçon de chemin en pointillé. En outre, des problèmes d'affichage multiéchelle peuvent s'ajouter.

Bien que ces 5A définissent un cadre de référence en termes de fonctionnalités, elles ne répondent pas aux spécificités des applications. Pour cela, l'une des approches actuelles consiste à pouvoir interfacer ce cadre générique à des

applications plus spécifiques. Cela s'effectue soit par des outils appropriés comme les outils d'analyse statistique soit par les langages de programmation.

3.1.2. *Parallèle entre les SGBD spatiaux et les SGBD relationnels*

Malgré leurs spécificités, les SGBD Spatiaux peuvent être vus comme des extensions des SGBD relationnels. Cette section établit un parallèle entre les SIG et les SGBD Relationnels (SGBDR).

	RELATIONNEL	SPATIAL
<i>Données</i>	Entier, Réel, Texte,...	Plus complexes : Point, Ligne, Région...
<i>Prédicats et calculs</i>	Tests : =, >, ... Calculs : +, /, ... et fonctions simples	Prédicats & calculs géom. et topo.: Tests : intersecte, adjacent à, ... Fonctions : intersection, surface...
<i>Manipulation</i>	Opérateurs de l'algèbre : Sélection, Projection, Jointure Agrégats: Count, Sum, Avg...	Manipulation mono ou inter-thèmes Sélection et jointure sur critère spatial Agrégats : fusion d'objets adjacents
<i>Liens entre objets</i>	Par clés de jointures	Relations spatiales (souvent) implicites
<i>Méthodes d'accès</i>	Index B-tree, hachage	Index R-tree, Quad-tree, Grid-file, etc.

Tableau 1. *Parallèle entre les SGBD spatiaux et les SGBD relationnels*

Ce parallèle est récapitulé par le tableau 1 qui montre à gauche les concepts bien connus dans le domaine des SGBDR [MAI 83] et à droite les concepts spécifiques aux SGBD spatiaux [BEN 91].

De haut en bas suivant le tableau 1, sur le premier point, les SGBDS ont développé de nouvelles représentations basées sur des données géométriques. Couramment dans un plan 2D, ce sont les lignes, les points et les régions. Ces nouvelles représentations constituent le prolongement des types de base des SGBDR comme les entiers, les réels, ou le texte.

Du côté opératoire, les simples prédicats de comparaison entre les valeurs alphanumériques dans un langage de requête comme SQL ne suffisent plus. Ces prédicats trouvent leurs équivalents en des opérateurs et prédicats pour capturer les relations spatiales pouvant exister entre les objets spatiaux [GUT 95]. Ces prédicats et d'autres opérateurs sont basés sur des algorithmes géométriques qui ont été

implémentés dans les SGBD spatiaux. Ces prédicats et opérateurs concernent, par exemple, l'intersection, l'adjacence, etc. qui sont décrits plus loin.

Du point de vue du langage de requête, les efforts se sont essentiellement concentrés sur des extensions du langage très populaire SQL. Ainsi, tous les opérateurs de bases comme la sélection, la jointure, trouvent un équivalent spatial. Plus particulièrement, pour l'opérateur fondamental de jointure dans les SGBDR, les algorithmes traditionnels trouvent leurs limites et ont été remplacés par de nouveaux algorithmes spécifiques [GUN 93]. Le développement de ces nouveaux opérateurs montre qu'un SIG n'est pas une simple extension d'un SGBDR. Enfin, un agrégat tel qu'une somme de valeurs numériques a un homologue spatial qui est la fusion d'objets adjacents.

Ces extensions ont rendu l'exécution de ces requêtes plus complexe. La complexité est due aux traitements qui utilisent le plus souvent l'algorithmique géométrique qui est très coûteuse en temps d'exécution. La complexité est due par ailleurs au volume des données qui est fréquemment d'une autre échelle de grandeur que dans les bases traditionnelles.

Parallèlement à ces extensions du langage, pour rendre efficace un SGBD spatial, des méthodes d'accès ont été développées. Ainsi, une méthode d'accès comme les *index B-tree* pour les données de dimension 1 (valeurs alphanumériques) trouvent une extension en des méthodes bien connues dans la communauté SIG comme les *R-Trees*. Le principe reste le même, mais l'application se fait sur des données géométriques rectangulaires. Pour l'autre grande famille de méthodes qu'est le hachage, celui-ci trouve son homologue dans des méthodes comme le *Grid-File*.

Enfin, d'un point de vue conceptuel, les liens sémantiques entre les données (caractéristique des SGBD relationnels) se traduisent par des relations spatiales. Différents modèles spatiaux ont cherché à capturer ces liens, soit explicitement par la représentation en un codage informatique des graphes (modèles topologiques), soit implicitement, par des calculs géométriques comme souligné précédemment.

En résumé, ces similitudes entre les SGBDR et les SGBD spatiaux ont permis de concevoir aujourd'hui des SIG puissants qui héritent des qualités des SGBDR.

3.1.3. Exemple de SIG existants

A l'heure actuelle, il existe de nombreux SIG sur le marché. Tous ne possèdent pas les mêmes caractéristiques, mais tous offrent une solution pour stocker les données traditionnelles. Leurs distinctions s'effectuent principalement sur la manière de gérer les données spatiales et leur architecture.

Le tableau 2 dresse un état des SIG les plus connus en France. Ces SIG gèrent les données sémantiques (factuelles), soit de manière spécifique (exemple *APIC*), soit en utilisant un SGBDR comme Oracle. Compte tenu du caractère précurseur de certains produits, le stockage des données spatiales a été fait pour la plupart dans des

systèmes conçus spécifiquement. La possibilité de stocker directement dans un SGBDR des données longues en binaire pour les valeurs spatiales existait dans le SGBD Empress-32 mais ne s'est étendue aux SGBD plus répandus que récemment. Dans Oracle, le stockage est possible depuis la version 7.

Produit	Données sémantiques	Données spatiales	Langage d'accès	Fournisseur
Apic	CCHR	CCHR	C/Fortran	APIC Système-FR
Arc/Info	Oracle Info	Arc	SML/AML	ESRI-USA
Argis	Oracle	Spécifique	C/Fortran/GQL	Unisys-FR
GDS	Oracle	Spécifique	GDS-Basic	GDS-UK
Geo/SQL	Oracle	AutoCAD	GeoSQL	Prosys-USA
GeoCity	Sybase	Spécifique	GeoLAG	Clemessy-FR
MGE	Oracle, Ingres	MicroStation	MDL	Intergraph-USA
SmallWorld	Oracle	Spécifique	Magik	Smallworld-USA
System 9	Empress-32	Empress-32	SQL géo	Prime Comp-USA

Tableau 2. *Les différents SIG du marché*

3.1.4. Architecture des SIG

Pour évaluer les différents SIG du marché, il est important de considérer leur architecture qui impacte leurs performances.

Il existe principalement (figure 3) trois grandes familles d'architecture pour la gestion des données sémantiques et des données spatiales [BEN 90b], [LAR 93]. Ces architectures influent sur les performances, la fiabilité et la cohérence dans la gestion des deux types de données. Ainsi, suivant la figure 3 :

– la gauche traduit la première génération de SIG qui place côte à côte un SGBD pour la gestion des données sémantiques et un système propre (appelé également boîte à outils) pour gérer les données spatiales. Ce type de couplage nécessite une couche d'intégration qui doit ventiler les demandes (requêtes) des applications vers l'une des deux parties.

- L'avantage est de pouvoir s'intégrer à n'importe quel SGBD du commerce dont il est impossible d'adapter le noyau en raison de l'inaccessibilité des codes sources. Le SGBD est alors un serveur du système qui est sollicité pour gérer ce qu'il sait faire au mieux, à savoir les données alphanumériques.

- Le désavantage est un couplage faible. Bien qu'un objet spatial soit constitué de données sémantiques et de données spatiales, le stockage dans les deux parties pose le problème de cohérence entre les deux systèmes. Par exemple, la restauration après une panne d'un côté ne garantit pas la cohérence avec l'autre côté. L'autre désavantage est de nécessiter que la couche au-dessus joue le rôle de médiateur pour synchroniser la gestion d'objets géographiques de part et d'autre. Ce rôle de médiateur se traduit par de nombreuses requêtes qui dégradent les performances globales du système.

Un exemple de système très connu et basé sur cette architecture est Arc-Info. C'est un précurseur qui offre néanmoins de nombreuses fonctionnalités et outils intéressants à l'heure actuelle.

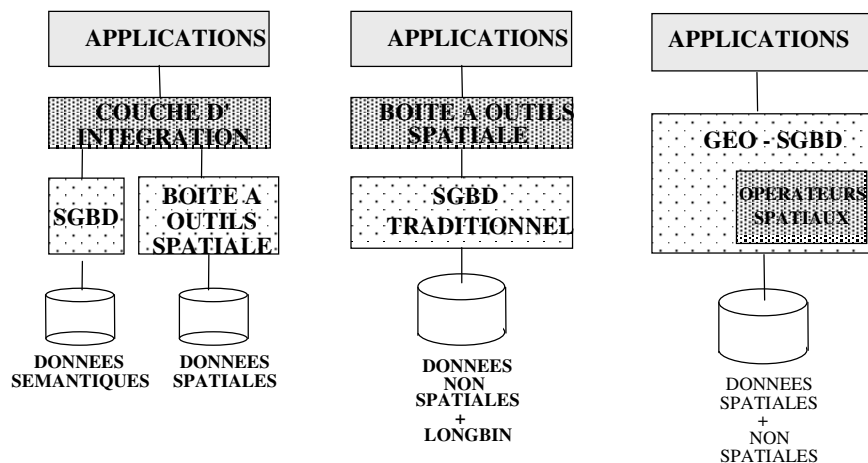


Figure 3. Les différentes architectures

- Le milieu de la figure montre une architecture qui résout en partie les problèmes cités précédemment. Les données spatiales et alphanumériques sont stockées uniformément dans le SGBD. Ceci a été rendu possible grâce aux nouvelles possibilités des SGBD de stocker des données binaires longues et non interprétées (appelé également BLOB⁴). La construction d'un SIG consiste alors à placer une surcouche chargée, comme dans le cas précédent, du traitement des opérations spatiales.

L'avantage est un stockage plus intégré des données spatiales. Cependant, l'exécution d'une requête spatiale reste peu performant car toute requête doit être interprétée par la couche supérieure qui doit ainsi solliciter de manière importante la

4. BLOB pour *Binary Long Object Bloc*. Le stockage dans une suite d'octets non interprétés, utilisé couramment pour le stockage d'images ou de sons.

couche du bas pour évaluer toute requête. Signalons que cette approche est récemment mise en œuvre dans Oracle 8 avec les notions de « cartridges ».

– La droite de la figure montre un système totalement intégré. Un tel système résout les problèmes évoqués. Cependant, l'offre est pauvre dans cette approche, car elle nécessite de concevoir un SGBD entier, ce qui requiert des ressources de développement très importantes. Il existe des prototypes de recherches dans cette catégorie d'architecture comme le SIG GéoSabrina [LAR 93].

En résumé, l'architecture peut influencer sur la fiabilité des données gérées et sur les performances. Les évolutions de ces architectures ont permis d'aller vers un couplage de plus en plus fort entre la partie sémantique et la partie spatiale. Ces évolutions garantissent des performances et une fiabilité accrues.

3.2. Modélisation de l'information géographique

L'information géographique est complexe. Sa modélisation s'appuie sur nombre de concepts, tels l'objet géographique, la couche thématique, les relations spatiales ou la méta-information spatiale. Cette section se propose de développer ces concepts.


3.2.1. Information géographique

L'information géographique se caractérise par une composante de localisation spatiale. Cette localisation s'inscrit sur la surface terrestre. En outre, on capture la morphologie des objets qui se traduit, entre autres, par la dimension de l'objet représenté, à savoir des points, des lignes ou des surfaces.

Une base de données géographique est définie par un ensemble d'objets géographiques organisé de manière à pouvoir être manipulé dans un SIG. L'information géographique capturée dans les systèmes actuels est souvent une projection au sol d'entités. On parle de coordonnées planimétriques.

Objet géographique ou spatial

L'objet géographique qui est la traduction d'une entité du monde réel, est représenté dans une base de données géographique par une structure. Cette structure contient à la fois des données sémantiques et des données spatiales (cf. figure ci-dessous). Les données alphanumériques décrivent qualitativement ou quantitativement des propriétés de l'objet. Par exemple, 3 millions d'habitants pour Paris. On parle de données « aspatiales » ou alphanumériques.

<577,'Paris', 3 733499,  >

Pour les données spatiales d'un objet géographique, elles s'appuient sur une représentation le plus souvent géométrique. La donnée géométrique peut ainsi décrire la morphologie de l'objet et sa localisation.

Couche thématique

Une couche thématique est un regroupement d'objets géographiques partageant les mêmes propriétés, les mêmes structures en une collection homogène. Ce regroupement définit un thème et est une approche naturelle de modélisation. Chaque thème peut ainsi désigner une couche. Par exemple, une couche peut représenter de l'hydrologie ou de la pédologie. Pour les traitements, on peut faire intervenir sélectivement les couches utiles. La superposition de ces couches suivant la métaphore des papiers calques constitue une carte. Du point de vue informatique, l'ensemble de ces couches thématiques constitue une base de données géographique.

Relations spatiales

Les relations spatiales sont des informations qui traduisent des propriétés essentielles dans le monde géographique. Tout géographe s'accorde à dire que tout phénomène en un endroit est lié à l'influence du voisinage et cette influence décroît avec l'éloignement (1^{ère} loi en géographie [TOB 79]).

Ces informations sur les relations spatiales peuvent exister explicitement ou implicitement dans une base de données géographique. Dans le cas explicite, on utilise des modèles de type topologique qui permettent de décrire ces relations à l'aide de graphes. Dans le cas implicite, les relations spatiales les plus simples sont déduites par des calculs géométriques. Ce calcul revient à effectuer le produit cartésien entre tous les couples d'objets spatiaux et à déterminer les couples qui répondent aux critères. Dans le cas d'une décomposition en couches thématiques, ces relations spatiales peuvent être intra ou inter thèmes.

Méta-informations géographiques

Pour exploiter les données d'une couche thématique, il est nécessaire d'avoir des méta-informations (information sur les données contenues). Ce sont typiquement les informations d'échelle, d'emprise, de référentiel géographique (système de projection), de qualité, de datation, ... Ces méta-informations s'appliquent le plus souvent sur un lot d'objets géographiques dit « lot homogène ». Ces informations servent à différents niveaux d'un SIG pour l'exploitation des données, soit par le système, soit pour renseigner l'utilisateur. En général, elles sont gérées dans un dictionnaire de données contenu dans le SIG.

3.2.2. *La représentation de données spatiales*

Cette section présente différentes techniques de représentation de l'information spatiale. Une carte qui contient une collection d'objets spatiaux n'est qu'une

abstraction ou une représentation symbolique du monde réel. En cela, une carte est limitée à un niveau de détail (et de précision). Cette limite est subjective. En outre, suivant le type de représentation spatiale utilisé, une dégradation de la représentation peut exister.

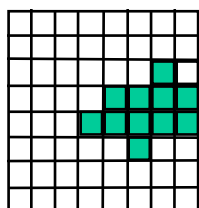


Figure 4. *Maillage régulier*

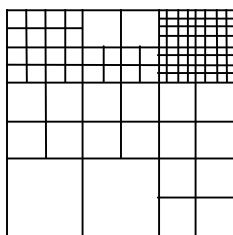


Figure 5. *Maillage irrégulier*

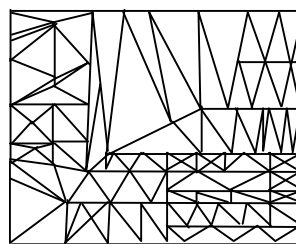


Figure 6. *Maillage irrégulier. Triangulation de Delaunay*

Il existe deux grandes perceptions pour la représentation des données. La première est une perception par étendu dont la représentation utilise des modèles basés sur la tessellation (ou le maillage) et la seconde par objet dont la représentation utilise le plus souvent des modèles vecteurs comme la représentation « spaghetti ».

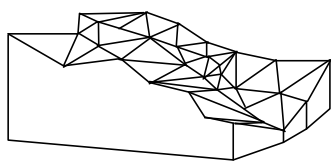


Figure 7. *Modèle numérique de terrain*

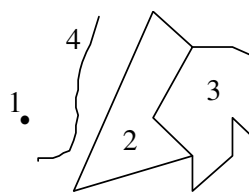


Figure 8. *Exemple de spaghetti*

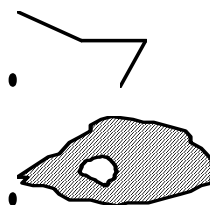


Figure 9. *Types spatiaux de base*

La tessellation consiste à décomposer l'objet de l'intérieur [SAM 89]. Cette décomposition peut aboutir à des mailles régulières (figure 4). L'exemple le plus courant concerne les images (appelées *rasters*) représentées par une matrice de pixels. D'un autre côté, la décomposition peut aboutir à des éléments irréguliers (figures 5 et 6). L'algorithme de Delaunay bien connu suivant le principe de triangulation donne ce type d'objets. Un autre exemple est illustré par les modèles numériques de terrains (figure 7).

Le modèle par tessellation présente plusieurs avantages. C'est une bonne représentation visuelle. Il permet d'effectuer à moindre coût les opérations d'interpolation, de seuillage ou d'agrégation. Ce sont des techniques similaires et

connues dans le domaine des traitements d'images. Un autre avantage concerne les sources de données du domaine géographique qui sont souvent sous format d'images satellites.

La seconde catégorie représente les données par leurs contours. Il existe principalement trois types de formats : spaghetti, réseau et topologique. Ces formats sont connus dans le monde géographique et dans les divers standards d'échange : EDIGEO, DIGEST, DLG, TIGER, ...

Le format spaghetti (figure 8) décrit les contours sans relations de contiguïtés et sans relations topologiques entre les objets. Par exemple, si l'on représente le contour d'une commune par une géométrie, le modèle n'indique pas directement quelles autres géométries lui sont contiguës. Le modèle réseau décrit la connexité des lignes. Ce modèle est principalement utilisé pour effectuer des traitements de types parcours de graphes. Un modèle plus riche est le modèle topologique qui modélise les contiguïtés entre toutes les formes d'objets, que ce soit des lignes, des nœuds ou des surfaces.

Pour leurs représentations géométriques, ces modèles utilisent l'un ou les trois types de base illustrés sur la figure 9. Ainsi, tout objet spatial peut être représenté par le type point, ligne ou région avec ou sans trous. Un trou représente une exclusion de surface. L'exemple typique est un lac dans une forêt. Ces trois types de base s'appuient sur les primitives géométriques tels que le point, la polyligne et le polygone. Ces primitives géométriques sont définies implicitement par un ou une liste de points décrivant leur contour. En outre, une fonction d'interpolation permet de relier ces points. La fonction couramment utilisée est une interpolation linéaire, mais cela peut être une fonction plus complexe comme une *spline*.

3.2.3. Codage des relations spatiales

Les langages d'extraction de données dans les SGBD spatiaux utilisent des opérateurs qui mettent en évidence des relations spatiales. Cette sous-section décrit ces relations pour ensuite exposer une technique de représentation de ces relations connue sous le nom d'indices de jointure. Cette technique s'apparente aux matrices de contiguïtés connues dans le domaine de l'analyse spatiale.

Catégories de relations spatiales

Les relations spatiales entre les objets spatiaux se traduisent dans un langage de requêtes par des prédicats (fonctions renvoyant une valeur binaire). De nombreuses études ont été faites pour classifier les relations spatiales pouvant exister. Le tableau 4 illustre les prédicats spatiaux les plus courants [EGE 93].

Prédicats binaires	Région	Ligne
Région	Adjacent, inclusion chevauchement	Borde chevauchement
Ligne	Gauche, droite, inclusion chevauchement	Connecte, inclusion chevauchement
Point	Inclusion	Extrémité, inclusion

Tableau 3. Exemples de relations spatiales

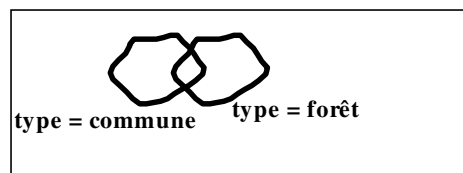


Figure 10. Requête graphique

Utilisation de ces prédicats

```
select  c.nom, f.nom
from    commune c, foret f
where   chevauchement (c.loc, f.loc)
```

Un exemple d'utilisation de ces prédicats est illustré par la requête ci-dessus. Cette simple requête sélectionne les noms des forêts et les noms des communes qui possèdent une forêt. Cette requête est une extension du langage SQL dans laquelle la clause WHERE a été étendue pour permettre d'utiliser des prédicats spatiaux comme ceux définis sur le tableau 3.

D'autres approches utilisent implicitement ces prédicats comme dans le cas de langage de requêtes graphiques [AUF 92]. Un exemple de formulation est illustré sur la figure 10. L'utilisateur dessine ce qu'il souhaite. L'ordinateur le traduit en une séquence d'opérateurs dont des prédicats du tableau 3 pour l'exécution de la requête.

Indices de jointure vs. Matrices de contiguïtés

Du point de vue du traitement, lorsque les relations spatiales entre les objets géographiques ne sont pas explicites, le coût de traitement pour calculer certaines relations peut être exorbitant. En effet, le calcul de ces relations revient souvent à effectuer le produit cartésien de deux collections d'objets spatiaux et à tester couple par couple afin de déterminer ceux qui répondent au critère. Une telle approche a un coût en $O(n^2)$. Autrement dit, le nombre d'opérations augmente comme le carré du

nombre d'éléments. Une solution pour réduire ce coût consiste à utiliser les indices de jointures.

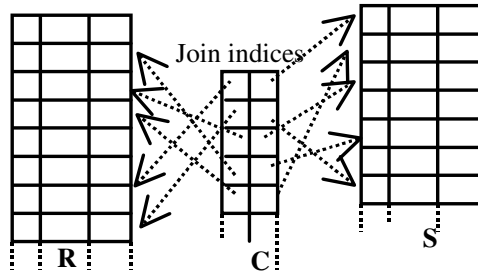


Figure 11. Un index de jointure

Obj1	Obj2	Distance
SR-1	PI-9	2.34
SR-1	PI-1	3.45
SR-2	PI-11	7.23
SR-2	PI-13	3.22
SR-2	PI-14	5.34
SR-3	PI-18	3.44
SR-3	PI-16	3.68
...		

Figure 12. Exemple matrice de contiguïté pondérée

La structure d'index de jointure a été proposée par Valduriez [VAL 87] comme une technique pour accélérer les jointures en relationnel. Le principe de cette technique consiste à utiliser une structure de type tableau (illustrée par C sur la figure 11) qui stocke des couples d'indices. L'indice de gauche référence un objet spatial d'une première collection d'objets (R sur la figure), et symétriquement, l'indice de droite référence un objet spatial d'une autre collection (S sur la figure). Les deux collections peuvent être identiques.

Cette technique traduit dans chaque couple d'indices l'existence d'une relation spatiale comme l'adjacence entre un couple d'objets spatiaux. L'extension de cette structure par le stockage d'une troisième colonne représentant la distance entre objets (figure 12) est particulièrement adaptée aux bases de données spatiales. En effet, un simple parcours de la collection C permet de connaître toutes les relations spatiales existant entre collections d'objets spatiaux [HJA 98]. En outre, le stockage de cette structure peut utiliser simplement les tables d'un SGBDR. Dans le cas spatial, cette technique a un intérêt accru en raison des calculs géométriques complexes et coûteux en temps. L'autre intérêt est de transformer un calcul de coût en $O(n^2)$ en un coût linéaire [DOR 91].

Une matrice de contiguïté a une structure très similaire. Elle code les relations spatiales par des couples d'indices. C'est une structure utilisée dans les problèmes d'analyse spatiale de données. Le critère peut être un critère de relations spatiales ou un critère de distance. Suivant la figure 12, l'exemple illustre tous les couples d'objets distants de moins de 8 m en précisant leurs distances.

4. Le Data Mining Spatial

Le DMS est une branche du Data Mining (DM). Afin de mieux situer ce vaste domaine, nous présentons brièvement la notion de Data Mining. Nous explicitons la terminologie utilisée en DM et ses principales fonctions. Cette section est un aperçu sur le DM. Pour des descriptions plus complètes, se référer à [FAY 96, GAR 99, LEF 98].

4.1. Le Data Mining

Le Data Mining (traduit en fouille de données) est né dans le contexte où des données de production se sont accumulées au fil du temps et où l'on s'est posé la question de leur devenir. Les alternatives étaient de ne rien en faire, de les archiver ou de trouver le moyen d'en extraire des informations cachées par analyse globale. C'est la troisième alternative qui a motivé le Data Mining. La métaphore est que cette accumulation de données (pas forcément produites dans un but d'analyse), peut délivrer, si on sait l'explorer (ou fouiller dedans) des informations utiles et précieuses.

L'objectif du DM est de découvrir des modèles (*patterns*) difficiles à mettre en évidence, soit en raison du volume important des données, soit à cause de la quantité de variables à considérer, soit enfin que ces modèles sont imprévisibles et ne sont jamais envisagés par l'analyste même à titre d'hypothèses à vérifier.

Le DM est couramment défini [FAY 96] comme l'extraction de connaissances intéressantes intelligibles (règles, régularités, patterns, contraintes) cachées dans les bases de données.

Pour [GAR 99] enfin, le DM consiste, depuis un ensemble de données, à découvrir des modèles : soit fonctionnels sous la forme $f(x_1, \dots, x_n) = y$ (par exemple : une régression linéaire $y = ax + b$), soit logiques comme les règles d'association ou les arbres de décision.

4.1.1. Domaines connexes au DM

Le DM s'inscrit dans un processus plus complet d'extraction de connaissances des données (ECD) ou Knowledge Discovery in Databases (KDD). Les étapes préalables au DM dans ce processus comprennent la construction d'un Data Warehouse et l'application de l'OLAP décrites ci-dessous. Ces deux dernières techniques sont définies dans [GAR 99], à savoir :

Un DW (*Data Warehouse*) ou un entrepôt de données est une base de données construite dans un but décisionnel depuis des bases de production souvent multisource. Un DW peut archiver des données historisées, actualisées de temps en temps, soit par interrogation des bases sources (data pull), soit par envoi automatique des modifications par les serveurs de données (data push). Une base dans

un DW est généralement de taille très importante à cause de l'archivage de données historisées au cours du temps.

L'OLAP (*On-Line Analytical Processing*) consiste en l'exploitation (en lecture) d'un DW par analyse multidimensionnelle et interactive. Il représente les données dans des « Data Cubes » donnant des comptages, totaux, etc., pour chaque variable et pour toute combinaison de variables avec différents niveaux de détail (par exemple: total annuel, sous-totaux mensuels, par semaine...).

4.1.2. Tâches génériques du DM

Les méthodes de DM peuvent être classées en deux catégories : les méthodes utilisées dans une phase exploratoire et les méthodes à caractère plus décisionnel qui cherchent à prédire une donnée (un attribut) particulière.

Les tâches relevant de l'analyse exploratoire sont de quatre types.

La description synthétique d'un ensemble d'objets : elle peut être basée sur les statistiques élémentaires lorsque l'analyse porte sur une à trois variable, au-delà elle rentre dans le cadre de l'analyse multidimensionnelle principalement basée sur l'analyse factorielle [LEB 97]. Une autre approche pour résumer les données est la généralisation conceptuelle des données qui simplifie les données en réduisant les détails sémantiques. Pour cela, [HAN 92] propose une méthode d'induction orientée attribut qui exploite des hiérarchies de concepts d'une base de connaissances.

La recherche de dépendances entre caractéristiques d'objets : l'analyse de correspondances est une des préoccupations de l'analyse de données multidimensionnelle. Elle permet de détecter si deux variables sont liées par l'Analyse Factorielle des Correspondances (AFC) ou entre les modalités des variables en appliquant l'Analyse de Correspondances Multiples (ACM). Une autre manière de découvrir les dépendances est la recherche de règles d'associations développée plus récemment [AGR 93, GAR 98] et appliquée à l'analyse de la consommation de produits commerciaux. Elle recherche des articles associés fréquemment dans une même transaction d'achat. Elle se base sur deux indices dits support s (fréquence de l'association) et confiance c (degré de vérification de la règle). Par exemple, on peut découvrir la règle :

télé → *magnétoscope* ($s=60\%$, $c=45\%$) où s = fréquence (*télé* & *magnétoscope*) et c = s /fréquence (*télé*)

En d'autres termes, si l'on achète une télévision, on a une forte probabilité d'acheter le magnétoscope également.

La classification automatique d'objets (dite clustering) : c'est une méthode d'apprentissage non supervisée qui, à partir de données non structurées, (*i.e.* fournies en vrac), décrit en extension une partition de ces données dans des classes. Elle se fonde sur une mesure de similarité (basée sur un critère de distance) dans le but de regrouper les données les plus similaires et de séparer les données éloignées. Hormis

la fonction distance, cette méthode ne demande aucune intervention de l'utilisateur. Elle est appliquée à l'ensemble des données et non à un échantillon.

Détection de tendances (trend) et de déviations : pour décrire la tendance en effaçant l'influence des données extrêmes ou atypiques, on utilise des mesures telles que la médiane au lieu de la moyenne. Les déviations sont détectées en utilisant des tests statistiques sur les écarts. Cette tendance peut être temporelle, comme par exemple l'analyse de l'évolution des stocks.

Dans l'analyse prédictive, plus orientée par l'utilisateur que la phase exploratoire, on distingue deux principales fonctions, à savoir :

Recherche de règles de classement d'objets : c'est une méthode d'apprentissage supervisé, qui, à partir d'une base d'exemples et d'un attribut à prédire, induit une description en intention permettant de classer les prochaines données. Le résultat est un arbre de décision, des règles ou un réseau de neurones. Contrairement à l'analyse exploratoire, la partie conclusion des règles induites est connue, ce sont les valeurs possibles de l'attribut à prédire. Une application possible concerne l'aide à la décision dans l'attribution de prêts aux futurs clients d'une banque en fonction des règles décelées dans la base de données de ses clients actuels.

Régression : cette méthode permet de prédire une variable en recherchant une fonction mathématique de type : $y = a_1*x_1 + a_2*x_2 + \dots + a_n*x_n + r$, avec r le résidu et y la variable à prédire. C'est donc un modèle fonctionnel. Un exemple est de calculer et de prédire un pourcentage de profit ou de perte des prêts selon un ensemble de variables quantitatives.

Suivant ces méthodes, on constate que le DM se base largement sur les acquis du domaine des statistiques, pour les approches numériques, et de l'intelligence artificielle pour les approches logiques [LEF 98]. Hormis la méthode de recherche d'associations, les travaux ont concerné essentiellement :

- l'amélioration des algorithmes pour les appliquer à de gros volumes de données en appliquant des techniques de bases de données comme l'indexation,
- l'intégration du DM aux SGBD et à SQL comme DMQL du système DBMiner [HAN 96],
- l'intégration du DM avec l'OLAP appelé OLAM [HAN 98a].

4.2. Principales approches du Data Mining Spatial

Le Data Mining spatial est défini comme l'extraction de connaissances implicites, de relations spatiales ou d'autres propriétés non explicitement stockées dans la base de données spatiales. Ses avantages sont, d'une part, son aspect exploratoire car, contrairement à l'analyse classique, il génère des hypothèses puis les valide et, d'autre part, il permet l'intégration complète de l'information sur la localisation spatiale et des liens de voisinage.

4.2.1. Spécificités du DMS

Comme il a été souligné dans la section 2.3., l'analyse de données spatiales nécessite l'analyse des interactions dans l'espace. Les méthodes de DM classique ne sont pas adaptées aux données spatiales car elles ne considèrent pas ces relations spatiales. Il est donc nécessaire de développer de nouvelles méthodes pour le Data Mining spatial et d'intégrer les techniques SIG et du DM.

Ces relations spatiales sont communément formalisées par la notion de graphe de voisinage, et peuvent utiliser une représentation sous forme de matrice de voisinages. Celle-ci est une matrice binaire M où $M[i, j]=1$ si l'objet i est voisin de l'objet j et $M[i, j]=0$ dans le cas inverse. Ceci est illustré par la figure 13.

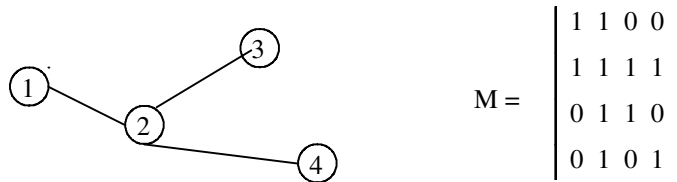


Figure 13. Graphe de voisinage et matrice de voisinage

La notion de voisinage est générale et peut aussi bien représenter une contiguïté entre formes zonales ou une proximité sur des points. Elle peut être étendue à une matrice de poids en qualifiant la proximité par une distance. Dans ce cas, elle n'est pas binaire. La matrice de voisinage est généralement symétrique, sauf si le graphe est orienté. Les distances par la route, en tenant compte des voies à sens unique, sont, par exemple, une illustration de graphes orientés.

4.2.2. Travaux sur le DMS

Les méthodes de DMS sont, pour la plupart, une extension de celles du Data Mining classique. Cependant, elles engendrent plus de problèmes de performances en raison du volume de données important généré par le codage des localisations géométriques et la complexité du calcul des relations spatiales. La recherche sur le DMS vise donc à proposer et à optimiser des méthodes d'analyse tenant compte des relations spatiales.

Comme ce domaine est à la croisée de plusieurs disciplines, cette recherche est menée au sein de deux communautés : des statisticiens s'intéressant à l'analyse spatiale et des chercheurs en bases de données [ZEI 00].

A l'heure actuelle, en matière de bases de données, on trouve essentiellement deux équipes : 1) l'équipe de J. Han à l'université Simon Fraser de Vancouver [URL 1]. Cette équipe [HAN 97] a développé le prototype GeoMiner comme une extension de DBMiner [HAN 96] ; 2) l'équipe de H.P. Kriegel de l'université de

Munich qui, d'une part, a proposé des algorithmes de *clustering* et, d'autre part, a développé d'autres méthodes de DMS (caractérisation, classification, tendances) basées sur une structure de graphe de voisinage [EST 97, EST 00b]. Citons d'autres travaux comme STING [WAN 97] de l'université de Californie qui est centré sur le *clustering* et l'expression de requêtes du type du langage SQL. Citons également l'extension du DW spatial par [BED 97] à l'université de Laval.

Quant à l'approche statistique [CRE 93], elle comprend d'anciens travaux comme sur l'autocorrélation spatiale [CLI 73] et la géostatistique [ISO 87] et des recherches plus récentes comprenant plusieurs écoles [LON 99] : Anselin soutient l'analyse spatiale exploratoire ESDA (Exploratory Spatial Data Analysis) et plus spécialement interactive ; Openshaw [URL 2] utilise le calcul intensif pour rechercher les clusters ; Lebart étend l'analyse multidimensionnelle aux contraintes de contiguïté [LEB 84, LEB 00].

4.3. Panorama des méthodes de DMS

Les tâches types de DMS s'inscrivent généralement comme un prolongement des tâches de DM intégrant les données et les critères spatiaux. Ainsi, une première phase exploratoire permet une description synthétique (indice d'autocorrélation globale, généralisation, densité, lissage), de découvrir les écarts donnant les spécificités locales (autocorrélation locale ou analyse factorielle locale) ou de chercher des regroupements de données (clusters). Cette première phase permet de guider la phase décisionnelle suivante, où l'on procède à une analyse plus fine afin d'expliquer les écarts ou de caractériser les groupes (caractérisation, règles de classement ou d'associations). Nous allons décrire ces différentes méthodes.

4.3.1. Description synthétique

Bien avant l'ère des SIG, des mesures du degré de dépendance aux voisins, dites d'autocorrélation spatiale globale ont été étudiées. Elles exploitent, hormis les attributs de l'objet, la matrice de voisinage définie ci-dessus. Dans le cas où les données sont corrélées, il faut les simplifier afin de faire apparaître une tendance générale et d'enlever le détail de leur variation. Pour cela, il existe différentes approches dont celles basées sur la densité, l'analyse multidimensionnelle lissée ou la généralisation.

Autocorrélation spatiale globale

L'autocorrélation spatiale permet de mesurer la ressemblance entre voisins, en utilisant la notion de graphe de voisinage. Elle est souvent utilisée comme technique exploratoire pour indiquer si une modélisation spatiale est nécessaire. Elle s'applique à des données quantitatives rattachées à des objets polygonaux formant un découpage (ou lattice, voir 3.2.2). Elle se décline en deux méthodes

complémentaires : l'indice de Moran (en 1948) et l'indice de Geary (1954). A titre d'exemple, nous détaillons ce dernier.

L'indice de Geary teste si la variabilité d'une variable entre voisins (donnée par la notion de variance locale) est significativement différente de celle attendue d'un modèle aléatoire (donnée par la variance).

Etant donnée une matrice de voisinage M , la variance locale d'une variable $X = \{x_1, x_2, \dots, x_n\}$ est définie comme suit :

$$V_{loc} = 1/2m \sum_i \sum_j M(i,j) (x_i - x_j)^2 \quad \text{où } m = \sum_i \sum_j M(i,j) \text{ pour } i, j = 1..n$$

L'indice de Geary est défini comme le rapport c de la variance locale V_{loc} et de la variance V .

$$c = V_{loc} / V$$

$$\text{soit : } c = [1/2m \sum_i \sum_j M(i,j) (x_i - x_j)^2] / [1/2n(n-1) \sum_i \sum_j (x_i - x_j)^2]$$

L'absence d'autocorrélation spatiale (indiquant que les valeurs de X sont indépendantes de la structure spatiale) se traduit par $c=1$. A l'inverse, une forte autocorrélation spatiale positive (indiquant une ressemblance des valeurs d'objets voisins) correspond à c tendant vers 0. Enfin, une autocorrélation fortement négative (indiquant une variation des valeurs d'objets voisins) donne un indice allant jusqu'à 2 ou 3.

Analyse multidimensionnelle lissée

Le besoin de prise en compte de l'organisation spatiale a conduit aux travaux de [LEB 84] et [BEN 90a] sur l'extension de l'analyse exploratoire de données multidimensionnelles. Ce type d'analyse comprend deux familles de méthodes : la classification (clustering) présentée plus loin et l'analyse factorielle dont le principe commun est d'assimiler le tableau N lignes et P colonnes de données à un ensemble de P points à N dimensions. L'analyse factorielle tend à visualiser dans le meilleur espace réduit ces points. Elle constitue un outil précieux pour l'expert en analyse de données. Elle comprend différentes méthodes dont les principales sont, d'un côté l'analyse factorielle en composantes principales (ACP) pour des tableaux de mesures de type individu/variable et de l'autre, l'analyse factorielle des correspondances (AFC) opérant sur des tableaux de contingence (croisement des modalités de deux variables).

L'extension des méthodes factorielles pour la prise en compte de la contiguïté est décrite dans [BEN 90a]. L'analyse multidimensionnelle lissée est obtenue de la manière suivante. Connaissant une matrice de voisinage, le tableau initial est lissé en remplaçant chaque valeur par le barycentre de ses voisins. L'application de l'ACP ou de l'AFC se fait ensuite sur le tableau ainsi modifié.

Densité

Cette méthode part de localisation ponctuelle d'un phénomène à analyser. La répartition et les concentrations de ces points ne sont pas toujours visibles à l'œil nu en raison du nombre élevé d'observations et des superpositions de localisations (tels que des accidents situés dans le même carrefour). La méthode de densité permet d'exhiber visuellement l'« intensité » du phénomène. Elle « généralise » les points en des cercles grâce au balayage de l'espace pour calculer les intensités comme le nombre de points des cercles [BAN 99]. Pour plus de détails sur cette méthode, voir [BAN 00] dans ce numéro.

Généralisation spatiale

Cette méthode est une extension aux données spatiales de la généralisation basée sur l'induction orientée attribut proposée dans [HAN 92]. Elle consiste à substituer les valeurs estimées trop détaillées par des valeurs moins détaillées jusqu'au niveau de détail souhaité, puis à agréger et compter les tuples identiques ainsi obtenus. Cette méthode permet de résumer les données et constitue une première étape pour induire des règles d'associations. Elle nécessite au préalable de disposer de la donnée de « hiérarchies de concepts » décrite ci-après.

Une hiérarchie de concepts est définie pour un attribut. Elle décrit le passage des concepts les plus spécifiques – correspondant aux valeurs de l'attribut dans la base de données – au concept plus général de niveau supérieur. De proche en proche, on passe du niveau le plus spécifique au niveau le plus général, n exemple sur les jours de la semaine. Pour un attribut de type spatial, cette hiérarchie est appelée *hiérarchie spatiale* et correspond à une relation spatiale d'inclusion entre objets. Le découpage administratif en pays, régions, département, communes, etc. est un exemple. La figure 14(b) donne un autre exemple. Pour un attribut non spatial, on parle de *hiérarchie thématique*.

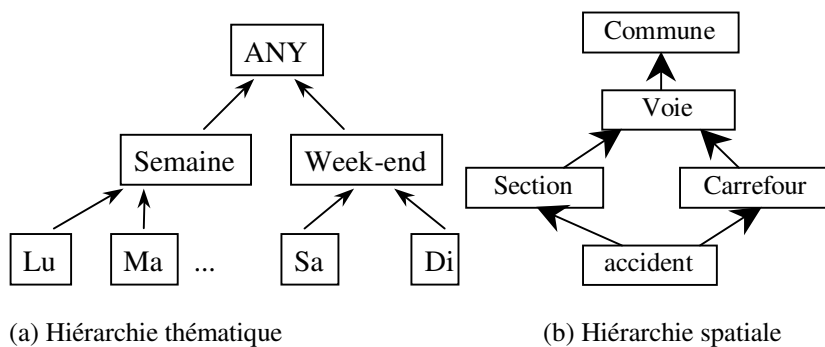


Figure 14. Hiérarchies de concepts

Deux types de généralisations spatiales ont été définis dans [LU 93] : La généralisation à dominante spatiale et la généralisation à dominante non spatiale.

La généralisation à dominante spatiale (GDS) exploite une hiérarchie spatiale existante en plus des hiérarchies thématiques. Pour illustrer la GDS, prenons l'exemple des hiérarchies de concepts ci-dessus avec une hiérarchie donnant la luminosité en fonction du moment de la journée. Partant de la carte des accidents, la GDS au seuil deux (c'est à dire que la substitution s'arrête lorsque les valeurs généralisées prennent au plus deux valeurs distinctes), engendre la remontée au niveau commune d'informations sur les accidents en les simplifiant et en faisant des comptages et des calculs (ici, la proportion des accidents par jour). Ceci est illustré dans la figure 15.

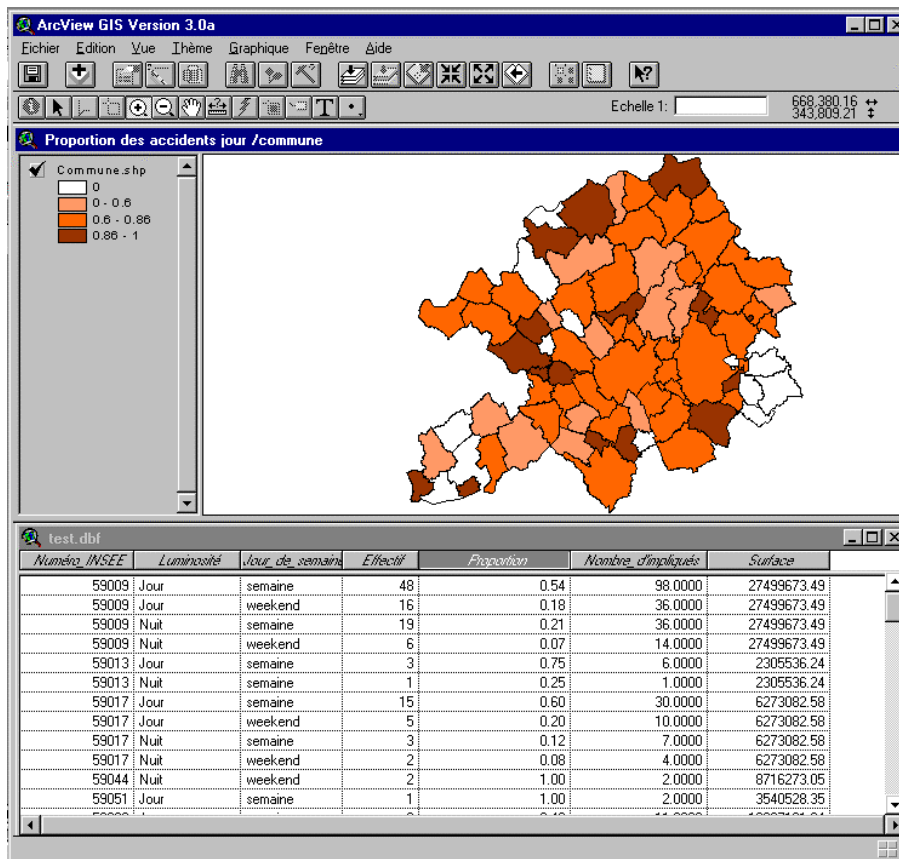


Figure 15. GDS d'accidents au niveau commune et selon le jour et la luminosité

La généralisation à dominante non spatiale (GDNS) n'utilise pas de hiérarchie spatiale mais génère des localisations moins détaillées par fusion d'objets spatiaux.

Une induction orientée attribut est faite en utilisant des hiérarchies thématiques, mais en gardant leur description spatiale. Cette induction produit des valeurs d'attributs homogènes (doublons) pour plusieurs objets. Ces objets sont alors fusionnés. Un exemple appliqué sur les accidents et utilisant les mêmes hiérarchies thématiques que l'exemple de GDS est donné figure 16. Ici, le résultat de la fusion de points donne un ensemble de points (dit multipoint). L'application de la GDNS à un découpage génère un nouveau découpage en zones plus grandes homogènes, pour le seuil demandé, vis-à-vis des attributs généralisés. [LU 93] donne l'exemple de zones avec un degré d'humidité dérivées de la carte des précipitations.

Cette méthode a été implémentée dans le prototype GeoMiner. Elle permet de générer des règles d'association sur les attributs généralisés qui ne sont pas déductibles au niveau de détail. Comme elle réduit le nombre de modalités des variables généralisées, elle peut également constituer une étape préalable à d'autres analyses telles que la recherche de règles caractéristiques [EST 98] ou des analyses de correspondances. En outre, ce mécanisme est à la base de l'extension au spatial de l'OLAP dans GeoMiner [HAN 98b] permettant d'explorer à différentes « échelles » sémantiques et spatiales les données.

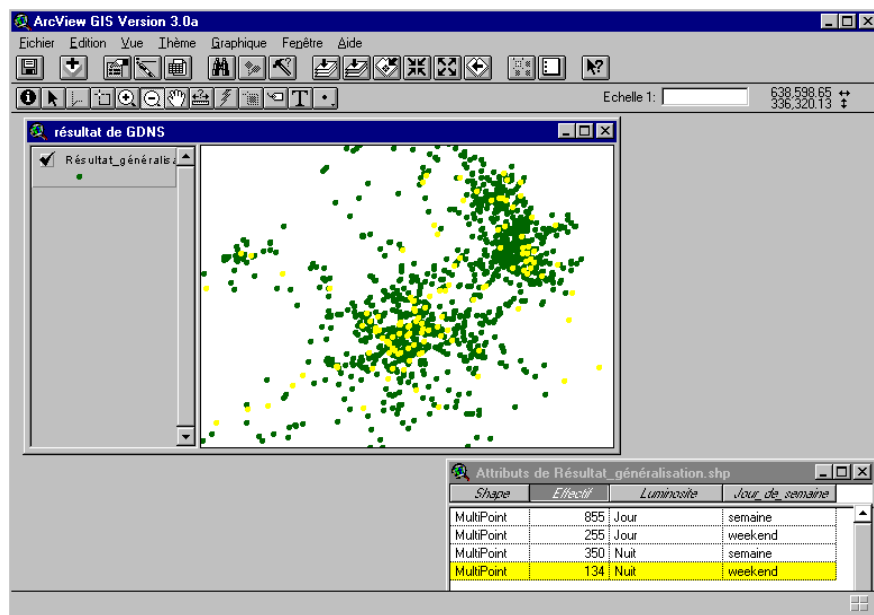


Figure 19. GDNS d'accidents selon le jour et la luminosité

4.3.2. *Analyse locale*

A l'inverse de l'analyse globale qui cherche à gommer les particularités, l'analyse locale vise à les ressortir pour mettre en évidence les données atypiques. C'est le cas de l'autocorrélation locale et de l'analyse multidimensionnelle locale.

Autocorrélation locale

Elle consiste à calculer un indice local affecté à chaque localisation d'objet. Cet indice mesure le degré de participation de l'objet à l'indice global pour une variable donnée. Il est dérivé de la formule de l'indice global où l'on remplace la matrice de voisinage par le vecteur ou la ligne de la matrice correspondant à l'objet. On parle aussi d'indicateur local d'association spatiale (LISA) [ANS 94].

Des exemples d'application en analyse de risque routier sont donnés dans l'article de Huguenin dans [HUG 00].

Analyse multidimensionnelle locale

Cette méthode est analogue à celle d'analyse lissée présentée ci-dessus. De la même manière, elle procède tout d'abord par transformation du tableau initial, mais en tableau contrasté. Ce dernier correspond à la différence du tableau initial et du tableau lissé présenté dans l'analyse locale. La suite est une analyse factorielle classique sur le tableau ainsi modifié.

4.3.3. *Clustering*

Le clustering est une méthode de classification automatique non supervisée qui regroupe des objets dans des classes. Son but est de maximiser la similarité intra-classes et de minimiser la similarité inter-classes. Elle est couramment utilisée en DM et est bien connue dans le domaine des statistiques [LEB 97]. Les principales sont celles par agrégation autour de centres mobiles, comme les k-means, les nuées dynamiques, puis la classification automatique hiérarchique (CAH). Le clustering permet, comme l'analyse factorielle, une analyse simultanée selon plusieurs variables.

La transposition au domaine spatial s'appuie sur une mesure de similarité d'objets localisés suivant leur distance métrique. Néanmoins, l'application de cette méthode au domaine spatial vise moins à classer qu'à détecter les endroits où il y a une concentration anormale (par exemple, détecter un point chaud dans l'étude de criminalité, ou des zones à risque en accidentologie). Cet écart est, soit absolu, c'est-à-dire comparé à un modèle aléatoire (clusters d'accidents de nuit), soit relatif, par rapport à une population de référence (clusters d'accidents de nuits non attendus par rapport à tous les accidents) [HUG 00].

Les travaux sur le clustering spatial sont surtout axés sur l'optimisation des algorithmes. Ainsi [EST 98b] propose des méthodes par partitionnement et hiérarchisation basées sur l'extension de DBSCAN en utilisant l'index spatial

R*tree, ainsi que des extensions incrémentales des algorithmes DBCLASS, DBLEARN.

Dans l'approche statistique, une fonction (dite *K*-fonction de Ripley [CRE 93]) permet de tester l'existence ou non de clusters, c'est-à-dire de déterminer si une distribution est aléatoire. Ce n'est pas la démarche de Openshaw qui prône l'utilisation de calculs intensifs pour résoudre le clustering. La machine GAM/K [URL 2] est une implémentation optimisée grâce à l'utilisation d'un index spatial.

Alors que le clustering à l'origine traite des tableaux à plusieurs variables, l'application aux données spatiales que nous venons de voir est basée uniquement sur la localisation (souvent sur un ensemble de points). L'extension aux attributs non spatiaux et aux objets de forme autre que ponctuelle a été proposée dans GDBSCAN en redéfinissant la fonction de similarité [SAN 98, EST 00a].

La classification automatique est aussi développée dans le domaine de traitement d'images comme pour l'interprétation d'images de télédétection. Elle sert à classifier les pixels des images. Elle utilise l'heuristique que des pixels d'une image proche ont plus de chance d'appartenir à la même classe. Notons que dans cette approche, le clustering n'est pas basé que sur la localisation, mais sur les variables en y intégrant le critère de proximité. Une méthode apparentée à cette approche est présentée dans [GOV 00].

Cette étape est souvent utilisée au préalable d'une autre tâche comme la recherche d'associations entre groupes et d'autres entités géographiques ou la caractérisation au sein d'un groupe.

4.3.4. *Analyse explicative*

Le terme explicatif est ici lié à une intervention de l'analyste qui, à la suite d'une découverte de clusters ou de valeurs atypiques par rapport à une tendance, focalise son analyse sur un sous-ensemble d'objets, sur une partie des variables, ou encore sur une zone géographique. Cette partie des données est ensuite analysée dans le but d'expliquer sa particularité par des liens avec certaines valeurs ou par des règles caractéristiques. Cette explication ne nécessite pas forcément l'utilisation de modèle au sens statistique.

Ces méthodes, à l'inverse des méthodes précédentes, opèrent sur plusieurs couches thématiques pour permettre d'expliquer un phénomène suivant les autres propriétés de son entourage. Ainsi, le risque d'accident peut être analysé en fonction des propriétés de l'accident et celles issues d'autres couches thématiques décrivant le tissu urbain (le bâti, la population, etc.). Ce qui permet d'« expliquer » l'occurrence d'accidents par la proximité à une classe d'objets géographiques (comme des centres commerciaux) qui sera découverte. Ici sont décrites les méthodes de caractérisation, de règles d'association et de classification.

Caractérisation

[EST 98a] définit la caractérisation comme l'induction des propriétés caractéristiques d'un sous-ensemble de données. Appliquée à des bases de données spatiales, la caractérisation découvre en plus le niveau d'extension de ces propriétés aux « voisins ». Une propriété caractéristique d'un sous-ensemble S est un prédicat $p_i=(attribut= valeur)$ tels que :

- sa fréquence relative dans S et dans son voisinage jusqu'à un ordre n est significativement différente par rapport à sa fréquence relative dans la base (rapport de fréquences supérieur à un seuil donné).
- sa fréquence relative est significativement différente dans le voisinage d'une proportion minimum d'objets du sous-ensemble S (proportion supérieure à un seuil de confiance).

Une description de cette méthode est également donnée dans [EST 00a]. Cette méthode a été appliquée à l'analyse du risque routier pour caractériser les accidents mortels par rapport à l'ensemble des accidents. Elle a été effectuée après une phase de généralisation sur les attributs permettant de réduire le nombre de propriétés à considérer et donnant une meilleure lisibilité du résultat. Elle a permis de déduire la règle suivante :

$$S \Rightarrow Cause = Cause\ humaine\ (0, 2.89) \wedge Cause = Indéterminée\ (0, 4.27) \wedge Jour\ de\ semaine = Weekend\ (0, 1.2) \wedge Luminosité = Nuit\ (0, 1.29)$$

Le module `Geo-Characterizer` de `GeoMiner` permet la découverte de règles caractéristiques à l'issue de la méthode de généralisation à dominante spatiale. Ces règles associent les données non spatiales avec leur localisation et ce pour chaque localisation.

Règles d'association

L'extension de la découverte de règles d'association de [AGR93] aux données spatiales permet de générer des règles de type :

$$X \rightarrow Y (s, c)$$

avec s comme support et c la confiance,

telles que X et Y sont des ensembles de prédicats spatiaux et non spatiaux. En d'autres termes, ceci revient à trouver des associations entre des propriétés des objets et celles de leurs « voisins ».

Le module `Geo-associator` de `GeoMiner` permet la découverte d'associations depuis un ensemble de couches thématiques donné, en utilisant certains attributs et des seuils de support et de confiance. Ainsi, la recherche d'associations impliquant les terrains de golf et les autres entités géographiques (bâtis, infrastructure, etc.) génère les deux règles suivantes :

$$is_a(x, "golf") \rightarrow close_to(x, "zone\ pavillonnaire") (61\%, 70\%)$$

$$is_a(x, "golf") \wedge close_to(x, "route\ secondaire") \rightarrow close_to(x, "park\ national") (50\%, 55\%)$$

L'algorithme proposé dans [KOP 95] propose deux phases dans l'évaluation du prédicat spatial. La première fait un test approximatif et génère des candidats pour un test précis du prédicat en seconde phase. Pour cela, des prédicats spatiaux généralisés (ou approchés) ont été introduits [KOP 00].

Classification

La recherche de règles de classement vise à structurer un ensemble d'objets en classes d'objets ayant des propriétés — selon une partie donnée des attributs-communes. Cette tâche est souvent réalisée par apprentissage supervisé qui, à partir de classes fournies partiellement en extension (un échantillon de la base de données), induit une description en intention permettant de classer les prochaines données. On parle de segmentation ou de *scoring* en statistique. La classification prend généralement la forme d'arbre de décision pour lequel l'algorithme de référence est ID3 [QUI 86].

L'extension au domaine spatial a été définie par [EST 97] par l'extension aux propriétés de leurs voisins jusqu'à un ordre N de voisinage. Ainsi, il est possible de trouver une règle de type :

Si population élevée et type de voisin = route et voisin de voisin = aéroport

Alors puissance économique élevée (à 95 %)

Une approche similaire a été proposée par [KOP 98] où, à la différence de la précédente, l'utilisateur précise les thèmes à explorer et l'ordre de voisinage est limité au premier (voisins directs).

5. Comparaison des approches au DMS

Cette étude a montré qu'il existe deux approches pour l'analyse et l'extraction de connaissances d'une base de données spatiales. La première est issue des statistiques spatiales et la seconde du domaine des bases de données (approche BD). Très peu de liens existent aujourd'hui entre ces deux types de recherches. Malgré cela, elles permettent parfois de résoudre les mêmes tâches d'analyse et ont certains points en commun. Cette section dresse une synthèse de ces similarités tout en soulignant les différences de ces approches. Elle montre aussi les forces et les faiblesses de chaque approche qui seront récapitulées dans le tableau ci-dessous.

5.1. Similarités

Pour les similarités, notons essentiellement l'utilisation et l'importance des relations de « voisinage » de toute sorte. En effet, les matrices de voisinage sont un préalable au calcul d'autocorrélation spatiale et les graphes des relations spatiales constituent une structure secondaire pour un bon nombre d'algorithmes en BD.

L'étape de calcul de ces relations est indispensable pour l'application de l'une ou l'autre de ces deux approches. Hormis la structure de matrice de voisinage, l'utilisation du critère de distance entre objets est très fréquent. C'est le cas du clustering, qui est une fonction essentielle de DMS. Des méthodes d'accès spatiales comme des index spatiaux sont nécessaires à l'optimisation d'une telle fonction.

5.2. Différences

L'approche BD différencie clairement les relations spatiales comme si c'étaient des propriétés à part entière. De plus, l'analyse statistique est plutôt orientée intra-thème, c'est-à-dire entre des individus d'un même thème, alors qu'en BD, elle peut également être inter-thèmes en mettant en jeu des jointures spatiales.

<p>APPROCHE STATISTIQUE</p>	<ul style="list-style-type: none"> + base mathématique solide : mesures, indicateurs + visuelle report sur des cartes (ESDA) + de multiples possibilités (\supset acquis des proba/stat) délivre des infos précises et quantifiées à l'analyste – monocouche thématique comme dans l'autocorrélation ou l'analyse de distribution données uniquement ponctuelles ou zonales – pas ou peu d'attributs certaines méthodes n'utilisent d'attributs qu'en phase de préparation (ex : cluster) d'autres portent sur une seule mesure seule l'AD contiguë porte sur plusieurs attributs – plus difficile à interpréter aux néophytes en stat./AD – ne découvre pas explicitement des règles spatiales
<p>APPROCHE BASE DE DONNEES SPATIALES</p>	<ul style="list-style-type: none"> +multithème : exploite/découvre les relations spatiales tout type de relations (connexité, distance, \cap, ...) toute forme d'objets (points, surfaces, lignes) + multi-attribut tout type d'attribut (mesure, qualitatif) + basé sur les techniques BDS (jointures, requêtes) + facilement interprétable induit directement des règles \supset des relations spatiales + utilise des connaissances sémantiques d'experts hiérarchies de concepts, analyse multiniveau – moins d'interactivité avec la carte (sauf OLAP spatial) – validité, robustesse moins sûre et moins mesurable pas assez de modèles – résultat généré parfois en nombre et difficile à filtrer – effet boîte noire

Tableau 4. Comparaison des approches au DMS

Les méthodes d'apprentissage produisent souvent des règles qui sont plus faciles à interpréter que les méthodes statistiques. Les méthodes statistiques, quant à elles, sont plus visuelles et certaines interagissent avec la carte.

Cette comparaison montre que chaque approche a ses atouts. Par conséquent, la combinaison des deux approches de DMS serait profitable dans le processus d'analyse. Dans les similarités, nous avons souligné le rôle central des relations spatiales, quelle que soit l'approche. Ces relations sont du ressort du SGBD géographique. A la différence du DM qui, généralement, n'utilise les fonctions de bases de données qu'en phase de préparation de données pour l'analyse, les BDS sont essentielles au DMS pour résoudre ces relations spatiales au sein de certaines méthodes. Les techniques d'optimisation pour le calcul de ces relations spatiales, présentées dans la section 3.2, trouvent ici tout leur intérêt.

6. Bibliographie

- [AGR 93] AGRAWAL R., IMIELINSKI T. & SWAMI A., "Mining Association Rules between sets of items in large databases", *Proceedings of the ACM SIGMOD*, Washington, DC, May 1993, p. 207-216.
- [ANS 94] ANSELIN L., "Local indicators of spatial association – LISA", *Research Paper 9331*, Regional Research Institute, West Virginia University, Morgantown, WV, 1994.
- [AUF 92] PORTIER-AUFAURE M-A., CIGALES: un langage d'interrogation de Systèmes d'Informations Géographiques, Doctorat de l'université Pierre et Marie Curie (Paris VI) - 9 octobre 1992.
- [BAN 99] BANOS A., BOLOT J., "Représentation surfacique d'événements ponctuels discrets - Comparaison méthodologique à partir d'accidents de la route", Actes de Colloque, *Quatrièmes Rencontres de Théo Quant*, Besançon, 1999.
- [BAN 00] BANOS A., "Quelle implication de l'utilisateur dans une stratégie de Data Mining spatial ? Illustration à partir de l'appréhension spatio-temporelle des accidents de la route en milieu urbain", *Revue internationale de géomatique n° 4/99*.
- [BED 97] BÉDARD, Y., LAM, S., PROULX, M.J., CARON, P.Y., LÉTOURNEAU, F., "Data Warehousing for Spatial Data: Research Issues", *Proceedings of the International Symposium Geomatics in the Era of Radarsat (GER'97)*, Ottawa, May, 1997.
- [BEN 90a] BENALI H., ESCOPIER B., "Analyse factorielle lissée et analyse factorielle des différences locales, *Revue Statistique Appliquée*, 1990, XXXVIII (2), pp 55-76.
- [BEN 90b] BENNIS-ZEITOUNI K., DAVID B., MORIZE-QUILIO I., VIÉMONT Y., "Géotropics: Database Support Alternatives for Geographic Applications", *Proc. of the 4th Int. Symposium on Spatial Data Handling*, Zurich, Suisse, 1990.
- [BEN 91] BENNIS-ZEITOUNI K., Un Modèle de Représentation et d'Organisation Physique des Données Géographiques, Thèse de Doctorat de l'Université Paris VI, 8 juillet 1991.
- [CLI 73] CLIFF A.D., ORD J.K., *Spatial autocorrelation*, Pion, London, 1973.

- [CRE 93] CRESSIE, N. A. C., *Statistics for Spatial Data*, Ed. Wiley, New York, 1993.
- [DOR 91] DORON R., "Spatial join indices", *Proc. 7th Conf. Data Engineering*, Kobe, Japan, 1991, 500-509.
- [EGE 93] EGENHOFER M.J. and SHARMA J., "Topological Relations Between Regions in R^2 and Z^2 ", *Advance in Spatial Databases, 5th International Symposium, SSD'93* p. 316-331. Singapore, June 1993, Springer-Verlag.
- [EST 97] ESTER M., KRIEGEL H.-P., SANDER J., "Spatial Data Mining: A Database Approach", *Proc. 5th Symp. on Spatial Databases*, Berlin, Germany, 1997.
- [EST 98a] ESTER M., FROMMELT A., KRIEGEL H.-P., SANDER J., "Algorithms for Characterization and Trend Detection in Spatial Databases", *Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining*, New York, NY, 1998.
- [EST 98b] ESTER M., KRIEGEL H.-P., SANDER J., XU X., "Clustering for Mining in Large Spatial Databases", *Special Issue on Data Mining, KI-Journal*, ScienTec Publishing, No.1, 1998.
- [EST 00a] ESTER M., "Algorithms for Spatial Clustering and Characterization", Actes des Journées sur le *Data Mining spatial et l'analyse du risque*, 24-25 février 2000, Versailles.
- [EST 00b] ESTER M., "The Database Approach to Spatial Data Mining", Actes des Journées sur le *Data Mining spatial et l'analyse du risque*, 24-25 Février 2000, Versailles.
- [FAY 96] FAYYAD *et al.*, "Advances in Knowledge Discovery and Data Mining", *AAAI Press / MIT Press*, 1996.
- [GAR 98] GARDARIN G., PUCHERAL Ph., WU F., "Bitmap Based Algorithms For Mining Association Rules", *Proceeding of European Conference BDA*, Tunis, October 1998.
- [GAR 99] GARDARIN G., *Internet / Intranet et bases de données : Data Web, Data Media, Data Warehouse, Data Mining*, Editions Eyrolles, 1999.
- [GOV 00] GOVAERT G., "Classification automatique et modèle de mélange : application aux données spatiales", *Revue internationale de géomatique n° 4/99*.
- [GUN 93] GÜNTHER O., "Efficient Computation of Spatial Joins", In *Proceeding of Data Engineering*, April 19-23, 1993, Vienna, Austria, p. 50-59.
- [GUT 95] GÜTING R.H., SCHNEIDER M., "Realm Based Spatial Data Types: The ROSE Algebra", *VLDB Journal*, vol. 4, p. 100-143, 1995.
- [HAN 92] HAN J., CAI Y. & CERONE N., "Knowledge Discovery in Databases; An Attribute-Oriented Approach." *Proceedings of the 18th VLDB Conference*, Vancouver, B.C., August 1992. p. 547-559.
- [HAN 96] HAN J., FU Y., WANG W., CHIANG J., GONG W., KOPERSKI K., LI D., LU Y., RAJAN A., STEFANOVIC N., XIA B., ZAIAANE O.R., "DBMiner: A System for Mining Knowledge in Large Relational Databases", *Proc. 1996 Int. Conf. on Data Mining and Knowledge Discovery (KDD'96)*, Portland, Oregon, August 1996, p. 250-255.
- [HAN 97] HAN J., KOPERSKI K., and STEFANOVIC N., "GeoMiner: A System Prototype for Spatial Data Mining", *Proc. ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'97)*, Tucson, Arizona, May 1997.

- [HAN 98a] HAN J., "Towards On-Line Analytical Mining in Large Databases", *ACM SIGMOD Record*, Vol. 27, p. 97-107, 1998.
- [HAN98b] HAN J., STEFANOVIC N., and KOPERSKI K., "Selective Materialization: An Efficient Method for Spatial Data Cube Construction", In *Proc. 1998 Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'98)*, Melbourne, Australia, April 1998.
- [HJA98] HJALTASON G.R. and SAMET H., "Incremental distance Join Algorithms for Spatial DataBases", *Sigmod 98*, Seattle, USA, p. 237-247.
- [HUG 00] HUGUENIN-RICHARD F., "Identifier les sites routiers dangereux – Application de méthodes d'analyse spatiale utilisant la localisation géographique des accidents", *Revue internationale de géomatique n° 4/99*.
- [ISO 87] ISOBEL C., "Practical geostatistics", *Applied Science Publisher*, Reprinted 1987.
- [JOS 00] JOSSELIN D., "A la recherche d'objets géographiques composites avec le prototype ARPEGE'", *Revue internationale de géomatique n° 4/99*.
- [KOP 95] KOPERSKI K. and HAN J., "Discovery of Spatial Association Rules in Geographic Information Databases", In *Advances in Spatial Databases (SSD'95)*, p. 47-66, Portland, ME, August 1995.
- [KOP 98] KOPERSKI K., HAN J., and STEFANOVIC N., "An Efficient Two-Step Method for Classification of Spatial Data", In *Proc. International Symposium on Spatial Data Handling (SDH'98)*, p. 45-54, Vancouver, Canada, July 1998.
- [KOP 00] KOPERSKI K., HAN J., MARCHISIO G. B., "Mining Spatial and Image Data through Progressive Refinement Methods", *Revue internationale de géomatique n° 4/99*.
- [LAR 93] LARUE T., PASTRE D., VIEMONT Y., "Strong Integration of Spatial Domains and Operators in a Relational Database System", *Proceedings of the 3rd Symposium on Spatial Databases (SSD'93)*, Singapour, June 1993, Lecture Notes in Computer Science n°692, Springer Verlag (Ed.), p. 53-72.
- [LAU 94] LAURINI R., THOMPSON D., "Fundamentals of Spatial Information Systems", Academic Press, London, UK, 680 p, 3rd printing, 1994.
- [LEB 84] LEBART L., "Correspondence analysis of graph structure", *Bulletin technique du CESIA*, 2, 5-19, 1984.
- [LEB 97] LEBART L. *et al.*, *Statistique exploratoire multidimensionnelle*, Editions Dunod, Paris, 2° édition, 1997.
- [LEB 00] LEBART L., "Contiguity analysis and related techniques", *Actes des Journées sur le Data Mining spatial et l'analyse du risque*, 24-25 Février 2000, Versailles.
- [LEF 98] LEFÉBURE R., VENTURI G., *Le Data Mining*, Eyrolles, 1998.
- [LON 99] LONGLEY P. A., GOODCHILD M. F., MAGUIRE D. J., RHIND D. W., *Geographical Information Systems - Principles and Technical Issues*, John Wiley & Sons, Inc., Second Edition, 1999.
- [LU 93] LU W., HAN J. and OOI B. C., "Discovery of General Knowledge in Large Spatial Databases", in *Proc. of 1993 Far East Workshop on Geographic Information Systems (FEGIS'93)*, Singapore, June 1993, p. 275-289.

- [MAI 83] MAIER D., *The Theory of Relational Databases*, Computer Science Press, 1983.
- [MAT 98] MATHSOFT Inc., "S-Plus for ArcView GIS - Users Guide", Version 1.0, *Data Analysis Products Division*, Seattle, Washington, April 1998.
- [QUI 86] QUINLAN J.R., "Induction of Decision Trees." *Machine Learning*, v.1, 1986. p.81-106.
- [SAM 89] SAMET H., *Applications of Spatial Data Structures*, Addison-Wesley, 1989.
- [SAN 89] SANDERS L., "L'analyse statistique des données en géographie", *GIP Reclus*, 1989.
- [SAN 98] SANDER J., ESTER M., KRIEGEL H.-P., XU X., "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications, in: *Data Mining and Knowledge Discovery*", *An International Journal*, *Kluwer Academic Publishers*, Vol. 2, No. 2, 1998.
- [TOB 79] TOBLER W. R., "Cellular geography", In Gale S. Olsson G. (eds) *Phylosophy in Geography*, Dortrecht, Reidel, p.379-86, 1979.
- [URL 1] URL : <http://db.cs.sfu.ca/GeoMiner/>, Site du projet GeoMiner.
- [URL 2] URL: <http://www/ccg.leeds.ac.uk/smart/gam/gam.html> , Site de la machine GAM.
- [VAL 87] VALDURIEZ P., "Join indices", *ACM Trans. on Database Systems*, 12(2); 218-246, June 1987.
- [WAN 97] WANG W., YANG J., and MUNTZ R., STING : A Statistical Information Grid Approach to Spatial Data Mining, Technical Report CSD-97006, Computer Science Department, University of California, Los Angeles, February 1997.
- [ZEI 98] ZEITOUNI K., Etude de l'application du Data Mining à l'analyse spatiale du risque d'accidents routiers par l'exploration des bases de données en accidentologie, Rapport de contrat PRISM -INRETS, décembre 1998.
- [ZEI 00] ZEITOUNI K., "A Survey on Spatial Data Mining Methods Databases and Statistics Point of Views", à paraître dans *IRMA 2000, Information Resources Management Association International Conference*, Data Warehousing and Mining Track, 21-23 May, 2000, Anchorage, Alaska, USA.