

CHAPITRE PREMIER

FOUILLE DE DONNEES SPATIALES

Marie-Aude Aufaure¹, Laurent Yeh², Karine Zeitouni²

¹ Laboratoire d'Ingénierie des Systèmes d'Information - INSA de Lyon

Bâtiment 501 - F-69621 Villeurbanne Cedex

Marie-Aude.Aufaure@lisi.insa-lyon.fr

² Laboratoire Prism - Université de Versailles

45, avenue des Etats-Unis - F-78 035 Versailles Cedex

{Prénom.Nom@prism.uvsq.fr}

1. INTRODUCTION

La fouille de données (data mining) spatiales est un domaine de recherche en pleine expansion. Elle s'attache à effectuer une analyse dirigée par les données sur des bases de données spatiales [20, 22, 29]. Elle offre de nouvelles perspectives pour beaucoup d'applications à caractère décisionnel comme le géo-marketing ou l'analyse de risques d'accidents [32] et dont le volume de données ne cesse de croître. Cette tendance à développer de grandes bases de données s'explique par les progrès dans l'acquisition de données spatiales et dans le géo-codage.

La Fouille de Données Spatiales (FDS) constitue un domaine de recherche à part entière. En effet, sa spécificité par rapport à la fouille de données traditionnelles [9] est qu'elle prend en compte les relations spatiales entre objets. Néanmoins, les tâches dans un processus de FDS sont définies comme une extension par les critères spatiaux de celles utilisées dans la fouille de données [33]. Ces différentes tâches permettent : (1) de résumer les données, (2) de trouver des règles de classification, (3) de former des clusters d'objets similaires, (4) de trouver des associations et des dépendances pour caractériser les données, et (5) d'extraire des tendances et de détecter les déviations.

Les méthodes permettant de réaliser ces tâches sont issues du domaine des statistiques et de celui des bases de données. Le point commun de ces méthodes est l'exploitation des relations spatiales de voisinage. Ceci ajoute un niveau de complexité dans l'implémentation de ces méthodes.

2 Titre ouvrage

Ce chapitre présente un état de l'art sur la fouille de données spatiales. Il est organisé de la manière suivante: la section 2 souligne les spécificités de la FDS. La section 3 présente les travaux sur ce domaine avant d'aborder un panorama des méthodes dédiées à la fouille de données spatiales.

2. SPECIFICITES DE LA FOUILLE DE DONNEES SPATIALES

2.1. La fouille de données spatiales

La fouille de données spatiales est définie comme l'extraction de connaissances implicites, de relations spatiales ou d'autres propriétés non explicitement stockées dans la base de données spatiales [14]. Ses avantages sont d'une part son aspect exploratoire dans le sens où les hypothèses sont générées, contrairement à l'analyse classique. D'autre part, les informations sur la localisation spatiale et sur les liens de voisinage sont complètement intégrées.

Les méthodes mises en œuvre pour la fouille de données spatiales utilisent de manière intensive les relations spatiales. C'est ce qui distingue ces méthodes de celles appliquées dans le cas de données de type alphanumérique. Ces relations spatiales jouent donc un rôle primordial dans l'analyse de données spatiales et la découverte de connaissances.

2.2. Caractéristiques des données spatiales

Les relations spatiales traduisent une caractéristique essentielle du monde réel. Elles mettent en évidence les relations et l'influence du voisinage entre les entités spatiales. La notion de dépendance spatiale est endémique aux données géographiques et est connue comme la 1^{ère} loi en géographie [30]. Il est donc essentiel d'en tenir compte dans l'analyse et l'interprétation des données localisées.

Les indices d'auto-corrélation spatiale définis en analyse spatiale considèrent uniquement les interactions sur une couche thématique. Or, en règle générale, les couches thématiques sont fortement corrélées. Une carte des précipitations et de la densité de population sont, par exemple, corrélées. En effet, la densité de population dépend de la production agricole qui, elle-même, est liée aux précipitations. Par conséquent, nous distinguons deux types de relations spatiales : (1) celles qui lient les

objets d'une même classe (appelées intra-thème) et (2) celles qui sont associées à plusieurs classes (appelées inter-thèmes).

2.3. Matrice et graphe de voisinage

Les relations spatiales sont communément formalisées par la notion de graphe de voisinage, et peuvent être représentées sous forme d'une matrice de voisinage [21]. 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 suivante.

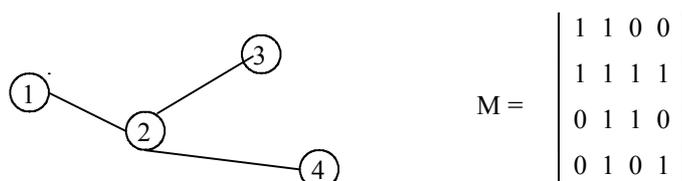


Figure 1 : Graphe de voisinage et matrice de voisinage

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 pondérée en qualifiant la proximité par une distance.

3. TRAVAUX SUR LA FOUILLE DE DONNEES SPATIALES

Bien que les méthodes dédiées à la FDS soient issues de celles de la fouille de données traditionnelles, elles engendrent plus de difficultés. Cela concerne des problèmes de performances en raison du volume de données spatiales et de la complexité du calcul des relations spatiales. Les travaux de recherche visent donc à proposer et à optimiser des méthodes d'analyse tenant compte des relations spatiales.

Ce domaine est à la croisée de plusieurs disciplines: les statistiques et les bases de données [34]. Une littérature abondante existe dans le domaine de l'analyse spatiale [10, 11], domaine largement antérieur à l'apparition des premiers SIG. Beaucoup de travaux portent sur la mesure de l'auto-corrélation locale et globale [5]. La géostatistique [24] se préoccupe d'analyser des tendances à l'aide de techniques de krigging. Des recherches récentes en analyse de données exploratoire et interactive

4 Titre ouvrage

ont été développées [2, 3, 18, 19]. Openshaw [25] prône l'utilisation du calcul parallèle pour identifier des clusters. Des travaux récents visent également à étendre l'analyse de données multidimensionnelle pour prendre en compte les contraintes de contiguïté [21].

Dans le domaine des bases de données, des algorithmes de fouille de données spatiales ont été proposés et des prototypes ont été développés. GeoMiner [12, 14] est une extension de DBMiner [13] et est couplé au SIG MapInfo. GeoMiner comme DBMiner ont l'avantage d'intégrer les techniques OLAP (On-Line Analytical Process) et fouille de données. Un moteur SQL étendu, appelé GMQL (GeoMiner Query Language), permet d'exécuter les opérations de fouille de données définies. Certains algorithmes sont basés sur une connaissance à priori exprimée par des hiérarchies de concepts.

Par ailleurs, les travaux de Ester se sont essentiellement articulés autour d'une structure de graphe de voisinage [6]. C'est le cas des méthodes comme la classification, la caractérisation et la recherche de tendances. Dans cette approche, les auteurs ont intégré au sein d'un SGBD géographique les opérateurs de graphe de voisinage. Citons d'autres travaux comme STING [31] centré sur le clustering et l'expression de requêtes dans un langage SQL étendu. Pour une bibliographie plus approfondie, le lecteur pourra se référer à [27].

4. PANORAMA DES METHODES DE FOUILLE DE DONNEES SPATIALES

Les méthodes types sont un prolongement des tâches de fouille de données intégrant les données et les critères spatiaux. Ainsi, une première *phase exploratoire* permet une description synthétique (indice d'auto corrélation globale, généralisation, densité, lissage), de découvrir les écarts donnant les spécificités locales (auto-corrélation locale ou analyse factorielle locale) ou de chercher des regroupements de données (clusters). Cette première phase permet de guider la *phase décisionnelle*, 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 maintenant décrire ces différentes méthodes et les illustrer avec des exemples issus des travaux de recherche actuels.

3.3.1 Phase exploratoire

Bien avant l'ère des SIG, des mesures du degré de dépendance aux voisins, dites d'auto corrélation spatiale globale ont été étudiées. Elles exploitent, hormis les attributs de l'objet, la matrice de voisinage définie au paragraphe 2.3. Ces mesures sont calculées à l'aide de deux méthodes complémentaires : l'indice de Moran (en 1948) et l'indice de Geary (1954). Dans le cas où les données seraient corrélées, il faut les simplifier afin de faire apparaître une tendance générale. Pour réaliser cela, il existe différentes approches dont celles basées sur la *densité*, l'*analyse multidimensionnelle lissée* ou la *généralisation*.

Généralisation spatiale

Cette méthode est une extension aux données spatiales de la généralisation basée sur l'induction orientée attribut proposée dans [12]. 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 n-uplets 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 d'une connaissance a priori que l'on trouve dans des « hiérarchies de concepts » définies par des experts. 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- aux concepts plus généraux de niveau supérieur. 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. en est un exemple. Pour un attribut non spatial, on parle de *hiérarchie thématique*.

La généralisation à dominante spatiale et la généralisation à dominante non spatiale ont été définies dans [23]. La généralisation à dominante spatiale exploite une hiérarchie spatiale existante en plus des hiérarchies thématiques. Les attributs sont générés par simplification et comptage. La généralisation à dominante non spatiale 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.

Cette méthode a été implémentée dans le prototype GeoMiner. Elle permet de générer des règles d'association sur les attributs généralisés qui

6 Titre ouvrage

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 [7] ou des analyses de correspondances. En outre, ce mécanisme est à la base de l'extension au spatial des techniques OLAP dans GeoMiner [15] permettant d'explorer les données à différentes échelles sémantiques et spatiales.

Analyse globale par lissage

L'**analyse multidimensionnelle lissée** est obtenue en remplaçant chaque valeur de la matrice de voisinage par le barycentre de ses voisins. L'application des méthodes d'analyse factorielle se fait ensuite sur le tableau ainsi modifié. L'extension des méthodes factorielles pour la prise en compte de la contiguïté est décrite dans [4].

Analyse locale

A l'inverse de l'analyse globale qui cherche à gommer les particularités, l'analyse locale vise à les faire ressortir pour mettre en évidence les données atypiques. Par exemple, si nous considérons les variables taux de scolarisation et revenu moyen, il n'existe a priori pas de corrélation au niveau global sur l'ensemble du territoire. Par contre, si l'analyse s'effectue au niveau local sur une région bien précise, une corrélation pourra être trouvée. L'analyse locale peut éventuellement contredire les résultats de l'analyse globale. L'auto corrélation locale consiste à calculer un indice local d'association spatiale [3], 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. L'analyse multidimensionnelle locale 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.

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 fouille de données et est bien

connue dans le domaine des statistiques [21].

La transposition au domaine spatial des méthodes de clustering s'appuie sur une mesure de similarité d'objets localisés suivant leur distance métrique. Néanmoins, la finalité du clustering en spatial n'est pas tant de former des classes que de détecter des concentration anormales (par exemple, détecter un point chaud dans l'étude de criminalité, ou des zones à risque en accidentologie). Les travaux sur le clustering spatial sont surtout axés sur l'optimisation des algorithmes. Ainsi [8] propose une extension de l'algorithme DBSCAN en utilisant l'index spatial R*tree, ainsi que des extensions incrémentales des algorithmes DBCLASS, DBLEARN. Openshaw prône l'utilisation de calculs intensifs et l'indexation spatiale dans sa machine GAM [25]. Alors que le clustering à l'origine traite des tableaux à plusieurs variables, l'application aux données spatiales 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 [28].

Cette étape est souvent utilisée en amont d'autres tâches de type décisionnelles comme la recherche d'associations entre groupes et d'autres entités géographiques ou la caractérisation au sein d'un groupe. Un exemple d'application est de former des clusters d'habitations puis de rechercher des caractéristiques communes par cluster.

3.3.2 Phase décisionnelle

Le terme explicatif ici est 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. Ces méthodes, à l'inverse des méthodes précédentes, opèrent sur plusieurs couches thématiques pour permettre d'expliquer un phénomène suivant les propriétés de son entourage. Nous décrivons les méthodes de caractérisation, de règles d'association et de classification.

Caractérisation

[7] définit la caractérisation comme l'induction des propriétés caractéristiques d'un sous-ensemble de données. Une règle caractéristique est une assertion qui décrit un concept satisfait par tous ou une grande

8 Titre ouvrage

partie des objets sélectionnés. 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 :

- (a) 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).
- (b) 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).

Le module Geo-Characterizer de GeoMiner permet la découverte de règles caractéristiques après application 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 [1] 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 leur voisinage.

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 un ensemble de règles comme par exemple :

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

L'algorithme proposé dans [16] 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 ont été introduits.

Classification

La recherche de règles de classement vise à structurer un ensemble

d'objets en classes d'objets ayant des propriétés communes. Cette tâche est 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 s'exprime généralement sous la forme d'un arbre de décision pour lequel l'algorithme de référence est ID3 [26]. L'extension au domaine spatial a été définie par l'extension aux propriétés de leurs voisins jusqu'à un ordre N de voisinage [6]. 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 [17] 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).

CONCLUSION

La fouille de données spatiales dérive de la fouille des données traditionnelles. Néanmoins, elle présente une spécificité importante par la prise en compte des relations spatiales. Les matrices et les graphes de voisinage constituent un pivot dans la conception d'opérateurs dédiés à la fouille de données spatiales. De plus la finalité de l'opérateur de clustering diffère car en spatial il est plus intéressant de rechercher des zones de concentrations plutôt que de former des classes. Une autre différence réside dans la complémentarité entre l'analyse globale et locale.

Excepté GeoMiner, on ne dénombre à l'heure actuelle pas de prototype opérationnel. Les travaux dans le domaine concerne plutôt l'implémentation d'algorithmes spécifiques comme le clustering. Toutefois, dans le cas du domaine spatial, les coûts algorithmiques sont prohibitifs en raison de la complexité des opérateurs géométriques et du volume des données.

On constate aujourd'hui de plus en plus d'applications ayant besoin d'intégrer la dimension spatiale dans une analyse exploratoire, ce qui amènera à un développement du domaine de la fouille de données spatiales. En perspective, il serait intéressant de rapprocher les travaux issus des domaines de l'analyse de données et des bases de données. Ainsi il serait possible compléter l'approche exploratoire pure par des modèles et des hypothèses préétablis, comme les modèles probabilistes.

REFERENCES

1. Agrawal R., Imielinski T. & Swami A., Mining Association Rules between sets of items in large databases. *Proceedings of the ACM SIGMOD*. Washington, DC, pp. 207-216 (1993).
2. Andrienko, N. and Andrienko, G.: Interactive Maps for Visual Data Exploration, *International Journal of Geographical Information Sciences* 13 (4), pp. 355-374 (1999). See also URL: <http://borneo.gmd.de/and/icavis>
3. Anselin, L.: Local indicators of spatial association - LISA. *Geographical Analysis*, 27, 2, pp. 93-115 (1995).
4. Benali H., Escofier B., Analyse factorielle lissée et analyse factorielle des différences locales, *Revue Statistique Appliquée*, XXXVIII (2), pp 55-76 (1990).
5. Cliff A.D., Ord J.K.,: *Spatial autocorrelation*, Pion, London (1973).
6. Ester, M., Kriegel, H.-P., Sander, J.: Spatial Data Mining: A Database Approach, *Proceedings of the 5th Symposium on Spatial Databases*, Berlin, Germany (1997).
7. 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).
8. 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).
9. Fayyad et al.: *Advances in Knowledge Discovery and Data Mining*, AAAI Press / MIT Press (1996).
10. Fisher, M. and Getis, A.: *Spatial analysis – spatial statistics, behavioural modelling and neurocomputing*, Berlin, Springer (1997).
11. Fotheringham, S. and Rogerson, P.: *Spatial Analysis and GIS*, Taylor and Francis (1995).
12. Han J., Cai Y. and Cerone N.: Knowledge Discovery in Databases; An Attribute-Oriented Approach., *Proceedings of the 18th VLDB Conference*, Vancouver, B.C (1992) pp. 547-559. See also URL : <http://www.cs.sfu.ca/~han>
13. Han J., Fu Y., Wang W., Chiang J., Gong W., Koperski K., Li D., Lu Y., Rajan A., Stefanovic N., Xia B., Zaiane 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 (1996) pp. 250-255.
14. Han J., Koperski K., and Stefanovic N.: GeoMiner: A System Prototype for Spatial Data Mining, *Proc. 1997 ACM-SIGMOD Int'l Conf. on Management of Data (SIGMOD'97)*, Tucson, Arizona (1997) System prototype demonstration.
15. Han J., Stefanovic N., and Koperski K., Selective Materialization: An Efficient Method for Spatial Data Cube Construction', *Proc. 1998 Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'98)* , Melbourne, Australia (1998).
16. Koperski K. and Han J., Discovery of Spatial Association Rules in Geographic Information Databases, *Advances in Spatial Databases (SSD'95)*, pp. 47-66, Portland, ME (1995).
17. Koperski K., Han J., and Stefanovic N., An Efficient Two-Step Method for Classification of Spatial Data', *Proc. International Symposium on Spatial Data Handling (SDH'98)* , pp. 45-54, Vancouver, Canada (1998).
18. Kraak, M.J. and MacEachren, A.M.: Visualisation for exploration of spatial data. *International Journal of Geographical Information Sciences* 13 (4), pp. 285-287 (1999).
19. Kraak, M.J.: Visualizing spatial distributions. Chapter 11 in Longley, P., M. Goodchild, D. Maguire & D. Rhind (editors) *Geographical information systems: principles*,

- techniques, management and applications*. New York: J. Wiley & Sons (1999) pp.157-173.
20. Laurini R., Thompson D.: *Fundamentals of Spatial Information Systems*, Academic Press, London, UK, 680 p, 3rd printing (1994)
 21. Lebart L. et al., *Statistique exploratoire multidimensionnelle*, Editions Dunod, Paris, 2^e édition (1997).
 22. Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W. : *Geographic Information Systems*, Volume 1, Wiley (1999).
 23. Lu W., Han J. and Ooi B. C., Discovery of General Knowledge in Large Spatial Databases', *Proc. of 1993 Far East Workshop on Geographic Information Systems (FEGIS'93)*, Singapore, pp. 275-289 (1993).
 24. Matheron, G. : Principles of geostatistics. *Economic Geology*, 58, pp. 1246-1266, (1963).
 25. Openshaw, S., Charlton, M., Wymers C. and Craft, A: A mark 1 geographical analysis machine for the automated analysis of point data sets, *International Journal of Geographical Information Systems*, Vol. 1 (4), pp. 335-358 (1987). See also URL : <http://www.ccg.leeds.ac.uk/smart/gam/gam.html>
 26. Quinlan J.R., Induction of Decision Trees, *Machine Learning*. vol.1, 1986. pp.81-106.
 27. Roddick, J.F, Spiliopoulou, M.: A Bibliography of Temporal, Spatial and Spatio-Temporal Data Mining Research, *ACM SIGKDD Explorations*, volume 1, Issue 1 (1999).
 28. 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).
 29. Scholl M. et al. *SGBD Géographiques, Spécificités*. International Thomson Publishing, Paris, France (1996) 185 p.
 30. Tobler W. R.: Cellular geography, In Gale S. Olsson G. (eds.) *Philosophy in Geography*, Dordrecht, Reidel (1979) 379-386.
 31. Wang, W., Yang, J. and Muntz, R.: STING+: An approach to active spatial data mining, *Proceedings of the Fifteenth International Conference on Data Engineering*, Sydney, Australia. (1999) IEEE Computer Society. pp. 116-12.
 32. 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* (1998).
 33. Zeitouni K., Yeh L., Les bases de données spatiales et le data mining spatial, *Revue internationale de géomatique*, Numéro spécial "Data mining spatial", Vol. 9, N° 4 (99). Egalement dans les actes des *Journées sur le Data Mining spatial et l'analyse du risque*, Versailles (2000).
 34. Zeitouni K., A Survey on Spatial Data Mining Methods Databases and Statistics Point of Views, *IRMA 2000, Information Resources Management Association International Conference, Data Warehousing and Mining Track*, Anchorage, Alaska, USA (2000).