A Survey of Spatial Data Mining Methods
Databases and Statistics Point of Views

Karine Zeitouni

PRiSM Laboratory - University of Versailles
45, avenue des Etats-Unis - F-78 035 Versailles Cedex

Tel / Fax : (0)1 39 25 40 46 / (0)1 39 25 40 57

Karine.Zeitouni@prism.uvsq.fr

ABSTRACT. This paper reviews the data mining methods that are combined with Geographic Information Systems (GIS) for carrying out spatial analysis of geographic data. We will first look at data mining functions as applied to such data and then highlight their specificity compared with their application to classical data. We will go on to describe the research that is currently going on in this area, pointing out that there are two approaches: the first comes from learning on spatial databases, while the second is based on spatial statistics. We will conclude by discussing the main differences between these two approaches and the elements they have in common.

KEYWORDS : Spatial Data Mining, Spatial Databases, Rules Induction, Spatial Statistics, Spatial Neighborhood.

1. INTRODUCTION

The growing production of maps is generating huge volumes of data that exceed people's capacity to analyze them. It thus seems appropriate to apply knowledge discovery methods like data mining to spatial data. This recent technology is an extension of the data mining applied to alphanumerical data on spatial data. The main difference is that spatial analysis must take into account spatial relations between objects.

The applications covered by spatial data mining are decisional ones, such as geomarketing, environmental studies, risk analysis, and so on. For example, in geomarketing, a store can establish its trade area, i.e. the spatial extent of its customers, and then analyze the profile of those customers on the basis of both their properties and the properties related to the area where they live.

In our project, spatial data mining is applied to traffic risk analysis [34]. The risk estimation is based on the information on the previous injury accidents, combined to thematic data relating to the road network, population, buildings, and so on. The project aims at identifying regions with a high level of risk and analyzing and explaining those risks with respect to the geographic neighborhood. Spatial data mining technology specifically allows for those neighborhood relationships.

Nowadays, data analysis in geography is essentially based on traditional statistics and multidimensional data analysis and does not take account of spatial data [31]. Yet the main specificity of geographic data is that observations located near to one another in space tend to share similar (or correlated) attribute values. This constitutes the fundamental of a distinct scientific area called “spatial statistics” which, unlike traditional statistics, supposes inter-dependence of nearby observations. An abundant bibliography exists in this area, including well-known geostatistics, recent developments in Exploratory Spatial Data Analysis (ESDA) by Anselin and Geographical Analysis Machine (GAM) by Openshaw. For a summary, refer to Part 1.c of [21]. Multi-dimensional analytical methods have been extended to support contiguity [19, 20]. We maintain that spatial statistics is a part of spatial data mining, since it provides data-driven analyses. Some of those methods are now implemented in operational GIS or analysis tools.

In the field of databases, two main teams have contributed to developing data mining for spatial data analysis. The first one, DB Research Lab (Simon Fraser University, Vancouver), developed GeoMiner [22], which is an extension of DBMiner. The second one (Munich University) devised a structure-of-neighborhood graph [6], on which some algorithms are based. They have also worked on a clustering method based on a hierarchical partitioning (extension of DBSCAN with a R*Tree), classification (extension of ID3 and DBLearn), association rules (based upon an efficient spatial join), characterization and spatial trends. STING (University of California) uses a hierarchical grid to perform optimization on the clustering algorithm [32]. We might also mention work on Datawarehouse dedicated to spatial data (University of Laval) [1].

This paper will describe data mining methods for Geographic Information Systems and highlight their value in performing spatial data analysis. It will survey both statistical approaches and those involving inference from databases.

It is structured as follows. In section 2 we define spatial data mining and subdivide it into generic tasks. Then in section 3 we classify spatial data mining methods, whether drawn from the realm of databases, statistics or artificial intelligence, in terms of these different tasks. We go on to compare the statistical analysis approach with the spatial database approach, with the aim of emphasizing their similarities and complementarity. Lastly, we conclude and discuss research issues.

2. DEFINITION OF SPATIAL DATA MINING

Spatial data mining (SDM) consists of extracting knowledge, spatial relationships and any other properties which are not explicitly stored in the database. SDM is used to find implicit regularities, relations between spatial data and/or non-spatial data.

The specificity of SDM lies in its interaction in space. In effect, a geographical database constitutes a spatio-temporal continuum in which properties concerning a particular place are generally linked and explained in terms of the properties of its neighborhood. We can thus see the great importance of spatial relationships in the analysis process. Temporal aspects for spatial data are also a central point but are rarely taken into account.

Data mining methods [8] are not suited to spatial data because they do not support location data nor the implicit relationships between objects. Hence, it is necessary to develop new methods including spatial relationships and spatial data handling. Calculating these spatial relationships is time consuming, and a huge volume of data is generated by encoding geometric location. Global performances will suffer from this complexity.

Using GIS, the user can query spatial data and perform simple analytical tasks using programs or queries. However, GIS are not designed to perform complex data analysis or knowledge discovery. They do not provide generic methods for carrying out analysis and inferring rules.

Nevertheless, it seems necessary to integrate these existing methods and to extend them by incorporating spatial data mining methods. GIS methods are crucial for data access, spatial joins and graphical map display. Conventional data mining can only generate knowledge about alphanumerical properties.

3. SPATIAL DATA MINING TASKS

As shown in the table below, spatial data mining tasks are generally an extension of data mining tasks in which spatial data and criteria are combined. These tasks aim to: (i) summarize data, (ii) find classification rules, (iii) make clusters of similar objects, (iv) find associations and dependencies to characterize data, and (v) detect deviations after looking for general trends. They are carried out using different methods, some of which are derived from statistics and others from the field of machine learning.

SDM Tasks	Statistics	Machine Learning
Summarization	Global autocorrelation Density analysis Smooth and contrast analysis Factorial analysis	Generalization Characteristic rules
Class identification	Spatial classification	Decision trees
Clustering	Point pattern analysis	Geometric clustering
Dependencies	Local autocorrelation Correspondence analysis	Association rules
Trends and deviations	Kriging	Trend rules

Table 1: Comparison between statistical and machine learning approaches to SDM

The rest of this section is devoted to describing data mining tasks that are dedicated to GIS.

3.1. Spatial data summarization

The main goal is to describe data in a global way, which can be done in several ways. One involves extending statistical methods such as variance or factorial analysis to spatial structures. Another entails applying the generalization method to spatial data.

3.1.1. Statistical analysis of contiguous objects

3.1.1.1. Global autocorrelation

The most common way of summarizing a dataset is to apply elementary statistics, such as the calculation of average, variance, etc., and graphic tools like histograms and pie charts. New methods have been developed for measuring neighborhood dependency at a global level, such as local variance and local covariance, spatial auto-correlation by Geary, and Moran indices [11, 24].

These methods are based on the notion of a contiguity matrix that represents the spatial relationships between objects. It should be noted that this contiguity can correspond to different spatial relationships, such as adjacency, a distance gap, and so on.

3.1.1.2. Density analysis

This method forms part of Exploratory Spatial Data Analysis (ESDA) which, contrary to the autocorrelation measure, does not require any knowledge about data. The idea is to estimate the density by computing the intensity of each small circle window on the space and then to visualize the point pattern. It could be described as a graphical method.

3.1.1.3. Smooth, contrast and factorial analysis

In density analysis, non-spatial properties are ignored. Geographic data analysis is usually concerned with both alphanumerical properties (called attributes) and spatial data. This requires two things: integrating spatial data with attributes in the analysis process, and using multidimensional data to analyze multiple attributes.

To integrate the spatial neighborhood into attributes, two techniques exist that modify attribute values using the contiguity matrix. The first technique performs a smoothing by replacing each attribute value by the average value of its neighbors. This highlights the general characteristics of the data. The other contrasts data by subtracting this average from each value.

Each attribute (called variable) in statistics can then be analyzed using conventional methods. However, when multiple attributes (above tree) have to be analyzed together, multidimensional data analysis methods (i.e. factorial analysis) become necessary [20]. Their principle is to reduce the number of variables by looking for the factorial axes where there is maximum spreading of data values. By projecting and visualizing the initial dataset on those axes, the correlation or dependencies between properties can be deduced.

In statistics and especially in the above methods, the analyzed objects were originally considered to be independent. The need to look at spatial organization spawned several research studies [20, 2]. The extension of factorial analysis methods to contiguous objects entails applying common Principal Component Analysis or Correspondence Analysis methods once the original table is transformed using smoothing or contrasting techniques.

3.1.2. Generalization

This method consists of raising the abstract level of non-spatial attributes and reducing the detail of geometric description by merging adjacent objects. It is derived from the concept of attribute-oriented induction as described in [22]. Here, a concept hierarchy can be spatial (like the hierarchy of administrative boundaries) or non-spatial (thematic) [13]. An example of thematic hierarchy in agriculture can be represented as follows: "cultivation type (food (cereals (maize, wheat, rice), vegetable, fruit, other)". That kind of hierarchy can be directly introduced by an expert in the field or generated by an inference process related to the attribute. A spatial hierarchy may preexist, like the administrative boundaries one, or it may be based on an artificial geometric splitting like a quad-tree [29], or it may result from a spatial clustering (see below).

There are two kinds of generalization: non-spatial dominant generalization, where we first use a thematic hierarchy and then merge adjacent objects; and spatial dominant generalization, which is based on a spatial hierarchy to begin with, followed by the aggregation or generalization of non-spatial values for each generalized spatial value. The complexity of the corresponding algorithms is O(NlogN), where N is the number of actual objects.

This approach could be treated as a first step towards a method of inferring rules, such as association rules or comparison rules.

3.1.3. Characteristic rules

The characterization of a selected part of the database has been defined in [5] as the description of properties that are typical for the part in question but not for the whole database. In the case of a spatial database, it takes account not only of the properties of objects, but also of the properties of their neighborhood up to a given level.

Consider a subset S of objects to analyze. This method uses the following parameters: 1) significance (relative frequency to the database in S); 2) confidence (ratio of objects in S which satisfy the significance threshold in the neighborhood) ; and 3) the maximum extension max-neighbors to the neighbors. This method throws up the properties *pi = (attribute, value*), the relative frequency factors *freq-fac i* (higher than the significance parameter) and the number *ni* of neighbors on which the frequency of the property is extended. The characterization can be expressed by the following rule:

$S \Rightarrow p\,1\,(n\,1\,,\,freq\text{-}fac\,1\,) \wedge \ldots \wedge p\,k\,(n\,k\,,\,freq\text{-}\,fac\,k\,).$

3.2. Class identification

This task, also called supervised classification, provides a logical description that yields the best partitioning of the database. Classification rules constitute a decision tree where each node contains a criterion on an attribute. The difference in spatial databases is that this criterion could be a spatial predicate and, because spatial objects are dependent on neighborhood, a rule involving the non-spatial properties of an object should be extended to neighborhood properties.

In spatial statistics, classification has essentially served to analyze remotely-sensed data, and aims to identify each pixel with a particular category. Homogeneous pixels are then aggregated in order to form a geographic entity [21].

In the spatial database approach [7], classification is seen as an arrangement of objects using both their properties (non-spatial values) and their neighbors' properties, not only for direct neighbors but also for the neighbors of neighbors and so on, up to degree N. Let us take as an example the classification of areas by their economic power. Classification rules are described as follows:

High population ^ neighbor = road ^ neighbor of neighbor = airport => high economic power (95%).

In GeoMiner, a classification criterion can also be related to a spatial attribute, in which case it reflects its inclusion in a wider zone. These zones could be determined by the algorithm, whether by clustering or by merging adjacent objects, or it could arise from a predefined spatial hierarchy.

A new algorithm [18] extends this classification method in GeoMiner to spatial predicates. For example, to determine high level wholesale profits, a decision factor can be the proximity to densely populated districts.

3.3. Clustering

This task is an automatic or unsupervised classification that yields a partition of a given dataset depending on a similarity function.

3.3.1. Database approach

Paradoxically, clustering methods for spatial databases do not appear to be very revolutionary compared with those applied to relational databases (automatic classification). The clustering is performed using a similarity function which was already classed as a semantic distance. Hence, in spatial databases it appears natural to use the Euclidean distance in order to group neighboring objects. Research studies have focused on the optimization of algorithms. Geometric clustering generates new classes, such as the location of houses in terms of residential areas. This stage is often performed before other data mining tasks, such as association detection between groups or other geographic entities, or characterization of a group.

GeoMiner combines geometric clustering applied to a point set distribution with generalization based on non-spatial attributes. For example, we may want to characterize groups of major cities in the United States and see how they are grouped. Cluster results will be represented by new areas, which correspond to the convex hull of a group of towns. A few points could stay outside clusters and represent noise. A description of each group may be generated for each attribute specified.

Many algorithms have been proposed for performing clustering, such as CLARANS [25], DBSCAN [6] or STING [32]. They usually focus on cost optimization. Recently, a method that is more specifically applicable to spatial data, GDBSCAN, was outlined in [15]. It applies to any spatial shape, not only to points data, and incorporates attributes data.

3.3.2. Statistic approach

Clustering arises from point pattern analysis [26, 9] and was mainly applied to epidemiological research. This is implemented in Openshaw's well-known Geographical Analysis Machine (GAM) and could be tested by using the K-function [4]. The clusters could also be detected by the ratio of two density estimates: one of the studied subset and the other of the whole reference dataset.

3.4. Spatial data dependencies

One way to reflect how data are related is the local autocorrelation method. The other typical for data mining yields association rules and has been adapted to spatial data.

3.4.1. Local autocorrelation

Local auto-correlation is concerned with the assessment of the degree of spatial dependence using the notion of spatial weight matrix [3, 27, 28]. This makes it possible to measure the difference between the actual spatial distribution of variable values and a random one. Thus, it is equivalent to a residual test in regression analysis. When the matrix consists of one column, association is sought between one point and all the others.

3.4.2. Association rules

This method is well known in data mining and is applied to market analysis by looking for items that are frequently associated in a commercial transaction [8]. It has been extended to deal with spatial data to express rules like:

A_1^A_2...^A_m ^ *Spatial Relations* => B_1^...^B_n ^ *Spatial Relations [s, c]* where Ai and Bj are predicates like attribute=constant_value , s is the rule support and c the rule confidence. These rules are used to find associations between properties of objects and those of neighboring objects.

For example, the rule :

is_a (x, gas_station) ^ within (x, rural_area) -> close_to (x, highway) [65%, 80%]

expresses the fact that gas stations that are located in rural areas are also close to highways at 80% and represent the majority (65%) of gas stations near highways.

Searching for association rules can involve the whole spatial database, as for example: “What kind of spatial objects are close to each other in California ?” with object types such as towns, forests, hydrology, roads, etc. The result is expressed by rules like:

is_a (x, big_town) ^ intersect (x, highway) → adjacent_to (x, river). [7%, 85%]

The main difficulty is to determine spatial relationships efficiently. The algorithm proposed in GeoMiner uses a concept of generalized predicates enabling spatial predicates to be evaluated in two phases [16, 17]. The first one performs an approximate test and generates candidates for performing an exact test of this spatial predicate in the second phase.

3.4.3. Extension to multi-level association rules

These associations can be generalized or detailed for forming a hierarchy of concepts. Spatial hierarchies or a conceptual hierarchy of attributes are refined or aggregated (like the subdivision into regions, then departments and municipalities). In relational databases, a current method entails aggregating classes of objects before looking for association rules in order to generate more general and relevant rules. An example of hierarchical association in a spatial database is to express the fact that 64 per cent of houses are about 500 meters from schools, two-thirds of which are primary schools and one-third secondary, or high, schools.

3.4.4. Group proximity rules

Suppose we have a cluster containing groups of private houses. The user wants to express the fact that the location of these groups is defined by the nearest particular spatial objects. For example, we can determine that 65 per cent of these houses are close to lakes, beaches or mountains.

[15] proposes a variation of association rules. The principle of this method is to discover the classes of objects which are frequently close to predefined groups. An algorithm CRH is used to make an efficient computation of the proximity of an object to a group (for example, aggregating the distances between this object and all the points in the group). Then another algorithm, called GenCom, is used to deduce the proximity rules and combine them with generalized attributes when a hierarchy of concepts is known.

3.5. Trend and Deviation Analysis

In relational databases, this analysis is applied to temporal sequences. In spatial databases, we want to find and characterize spatial trends.

3.5.1. Database approach

Using the process described in [7], which is based on the central places theory, the analysis is performed in four stages. The first one involves discovering centers by computing local maxima of particular attributes; in the second, the theoretical trend of these attributes is determined by moving away from the centers; the third stage determines the deviations in relation to these trends; and finally, we explain these trends by analyzing the properties of these zones. One example is the trend analysis of the unemployment rate in comparison with the distance to a metropolis like Munich. Another example is the trend analysis of the development of house construction.

3.5.2. Geostatistical approach

Geostatistics is a tool used for spatial analysis and for the prediction of spatio-temporal phenomena. It was first used for geological applications (the geo prefix comes from geology). Nowadays, geostatistics encompasses a class of techniques used to analyze and predict the unknown values of variables distributed in space and/or time. These values are supposed to be connected to the environment. The study of such a correlation is called structural analysis. The prediction of location values outside the sample is then performed by the “kriging” technique [14].

It is important to remember that geostastics is limited to point set analysis or polygonal subdivisions and deals with a unique variable or attributes. Under those conditions, it constitutes a good tool for spatial and spatio-temporal trend analysis.

4. COMPARISON OF SDM APPROACHES

One interest of this study is to bring together the whole body of research relating to the analysis and extraction of spatial data. The research was carried out either in the field of statistics, or in the field of database learning, but most of the time they ignored each other. One thus has to be able to compare and analyze them with the same analytical goal. After classifying them by task and distinguishing between the different methods arising from these two approaches, this section will seek to make a comparison of all these methods and identify the points they have in common. Here is a résumé:

4.1. Graphical methods and semantic methods

Some methods are based solely on the graphical aspect of the data, as in the exploratory analysis of spatial data (density and relative cluster). The result is often visual.

Others, on the other hand, utilize a semantic representation of spatial relations such as graphs and neighbor matrices. Apart from clustering, which remains a graphical method, most of the methods derived from the database approach fall into this category. In the statistical approach, one may describe auto-correlation tests, smoothing, and smoothed or contrasted factorial analysis as semantic methods.

4.2. Taking account of contiguity

There are substantial differences in the use of neighborhood semantics. In the learning approach spatial relationships are clearly represented, as though it were a question of properties in their own right. Conversely, in the statistical approach these neighborhood relationships are either integrated in formulas, as in the case of auto-correlation, or used to rectify the initial data, as in smoothed analysis.

Furthermore, in the statistical approach, these relationships are exclusively intra-thematic, which is to say among objects of the same theme, whereas they can also be inter-thematic (between several layers) in the learning approach. This is important, especially in an explanatory model where surrounding objects may intervene, whatever the theme. As an example, rainfall and population density layers are highly correlated.

Inter-thematic relationships are retrieved using joint operators with various spatial criteria. Since these operators are complex and time consuming, one needs to try and optimize them [12, 33].

4.3. Interpretation

In addition, the learning approach, like generalization, enables the data to be summarized and synthesized by aggregating them and combining their geographic locations. This approach generates classifications with very little intervention on the part of the user and produces association rules that non-specialists can understand.

Graphical methods forming part of exploratory analysis offer a very high degree of readability and require relatively little knowledge to use them.

As for factorial analysis, it also synthesizes the data, but, contrary to generalization, it does not reduce the number of objects, which may be a handicap for large amounts of data. The result may be of great interest for an enlightened user of these techniques who is capable of interpreting them, but not for a neophyte in data analysis.

4.4. Complementarity

These differences result in a degree of complementarity that is extremely valuable from an analytical viewpoint. For example, a generalization phase would enable the data to be reduced and simplified in order to prepare them for smoothed or contrasted factorial analysis.

It would also be interesting to undertake generalization prior to characterization, the search for associations or classification rules. Similarly, characterization or the search for associations may be used to explain a localized concentration.

Another approach is that described in [30]. It would entail carrying out a density analysis to find centers, then contrasting the real trend with a theoretical trend in order to detect deviations, and finally looking for properties that are characteristic of the places of these deviations.

5. CONCLUSION AND RESEARCH ISSUES

Different methods of data mining in spatial databases have been outlined in this paper, which has shown that these methods have been developed by two very separate research communities: the Statistics community and the Database community.

We have summarized and classified this research and compared the two approaches, emphasizing the particular utility of each method and the possible advantages of combining them. This work constitutes a first step towards a methodology incorporating the whole process of knowledge discovery in spatial databases and allowing the combination of the above data mining techniques.

Among the other issues in the area of spatial data mining, one approach is to consider the temporality of spatial data, while another is to see how linear or network shape (like roads) can have a particular influence on graphical methods. In any event, it remains essential to continue enhancing the performance of these techniques. One reason is the enormous volumes of data involved, another is the intensive use of spatial proximity relationships. In the case of graphical methods, these relationships could be optimized using spatial indexes. As regards the other methods that use neighborhood structures, instanciation of the structure is costly and should be pre-computed as far as possible.

ACKNOWLEDGEMENTS

This research forms part of a national PSIG project of the CASSINI network, dealing with the traffic risk analysis. My thanks to the participants in this project and especially to Sylvain Lassarre from INRETS (the French national institute for transport and safety research) and Florence Richard from the THEMA laboratory for their contribution to this study.

REFERENCES

1. 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, pp. 25-30
2. Benali, H., Escofier, B.: Analyse factorielle lissée et analyse factorielle des différences locales, Revue Statistique Appliquée, 1990, XXXVIII (2), pp 55-76
3. Cliff A.D., Ord J.K., 1973 : "Spatial autocorrelation", Pion, London.
4. Diggle P.J., 1993, Point process modeling in environmental epidemiology. In Barnett V., Turkman K. (eds) Statistics for the environment, Chichester, John Wiley & Sons, pp 89-110.
5. 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
6. Ester, M., Kriegel ,H.-P., Sander, J., Xu, X.: Density-Connected Sets and their Application for Trend Detection in Spatial Databases, Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining, Newport Beach, CA, 1997, pp. 10-15
7. Ester, M., Kriegel, H.-P., Sander, J.: Spatial Data Mining: A Database Approach, Proc. 5th Symp. on Spatial Databases, Berlin, Germany, 1997
8. Fayyad et al., "Advances in Knowledge Discovery and Data Mining", AAAI Press / MIT Press, 1996
9. Fotheringham S., Zhan B., 1996 : "A comparison of three exploratory methods for cluster detection in spatial point patterns", Geographical Analysis, Vol. 28, n° 3, pp. 200-218
10. Gatrell A., Bailey T., Diggle P., Rowlingson B., 1996 : "Spatial point pattern analysis and its application in geographical epidemiology", Transactions of the Institute of British Geographers, n° 21, pp. 256-274
11. Geary R.C.: The contiguity ratio and statistical mapping, The incorporated Statistician, 5 (3), pp 115-145.
12. Gunther O., "Efficient Computation of Spatial Joins", Proc of Data Engineering, Vienna, Austria, April 1990, pp. 50-59.
13. 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. pp. 547-559
14. Isobel C., "Practical geostatistics", Applied Science Publisher, Reprinted 1987. Also at URL: <http://curie.ej.jrc.it/faq/introduction.html>
15. Knorr E. M., and Ng R. T.: Finding Aggregate Proximity Relationships and Commonalities in Spatial Data Mining, IEEE Transactions in Knowledge and Data Engineering, Vol 8(6), December 1996.
16. Koperski K., " A Progressive Refinement Approach to Spatial Data Mining", PhD Thesis, the School of Computing Science, Simon Fraser University, April 1999.
17. Koperski, K. and Han, J.: Discovery of Spatial Association Rules in Geographic Information Databases, In Advances in Spatial Databases (SSD'95), pp. 47-66, Portland, ME, August 1995
18. Koperski, K., Han, J., and Stefanovi,c N.: An Efficient Two-Step Method for Classification of Spatial Data, In Proc. International Symposium on Spatial Data Handling (SDH'98) , pp. 45-54, Vancouver, Canada, July 1998
19. Lebart L. et al., "Statistique exploratoire multidimensionnelle" , Editions Dunod, Paris, 439 p., 1997.
20. Lebart, L. (1984) Correspondence analysis of graph structure. Bulletin technique du CESIA, Paris:2, 1-2, pp 5-19.
21. 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.
22. Lu, W., Han, J. and Ooi, B.: Discovery of General Knowledge in Large Spatial Databases, in Proc. of 1993 Far East Workshop on Geographic Information Systems (FEGIS'93), Singapore, June 1993, pp. 275-289
23. Mathsoft Inc., "S-Plus for ArcView GIS - Users Guide Version 1.0" and "S-Plus Spatial Stat.", Data Analysis Products Division, Seattle, Washington, April 1998.
24. Moran P.A.P., The interpretation of statistical maps, Journal of the Royal Statistical Society, B: 10, pp 234-251., 1948.

25. Ng, R. and Han, J.: Efficient and Effective Clustering Method for Spatial Data Mining, in Proc. of 1994 Int'l Conf. on Very Large Data Bases (VLDB'94), Santiago, Chile, September 1994, pp. 144-155

26. Openshaw S., Charlton M., Wymer C., Craft A., 1987 : "A mark 1 geographical analysis machine for the automated analysis of point data sets", International Journal of Geographical Information Systems, Vol. 1, n° 4, pp. 335-358

27. Ord J.K., Getis A., 1992 : "The Analysis of Spatial Association by Use of Distance Statistics", Geographical Analysis, Ohio Sate University Press, Vol. 24, n° 3, pp. 189-206

28. Ord J.K., Getis A., 1995 : "Local Spatial Autocorrelation Statistics : Distributional Issues and an Application, Geographical Analysis", Ohio State University Press, Vol. 27, n° 4, pp. 287-306

29. Samet H., "Design and Analysis of Spatial Data Structures: Hierarchical (quadtree and octree) data structures ", Addison-Wesley Edition, 1990

30. Sander, J., Ester M., Kriegel H.P., Xu X.: Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications, in: Data Mining an Knowledge Discovery", An International Journal, Kluwer Academic Publishers, Vol. 2 (2), 1998.

31. Sanders, L.: L'analyse statistique des données en géographie, GIP Reclus, 1989

32. 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

33. Yeh T. S.: Yeh T-S., "Spot: Distance based join indices for spatial data", ACM GIS 99, Kansass City, 5-6 November 1999.

34. 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, Final report of the contract PRISM -INRETS, December 1998, 33 p.