Spatial Data Mining Methods Databases and Statistics Point of Views

R. Regin¹, Dr S. Suman Rajest²*, Dr Bhopendra Singh³

¹Department of Information Technology, Adhiyamaan College of Engineering, Hosur, Tamil Nadu, India.
²Vels Institute of Science, Technology & Advanced Studies (VISTAS), Tamil Nadu, India.
³Amity University, Dubai.

¹regin12006@yahoo.co.in, ²sumanrajest414@gmail.com, ³bsingh@amityuniversity.ae

Abstract. This article reviews the approaches used in data mining to perform a geographical study of regional datasets coupled with Geographic Information Systems (GIS). Firstly, we can look at the functions of data mining used by such data and then illustrate their precision compared to their classic data use. We will further explain the research conducted in this sector and point out that two separate methods exist: one is focused on space database learning while the other is based on space statistics. Finally, we will address the key distinctions between these two methods and their similar features.

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

1. Introduction

The increasing development of maps generates massive quantities of data that surpass people's capacity to interpret them. Thus, techniques of exploration of information such as data mining seem suitable for spatial data. This latest technique is an expansion of alphanumeric data mining on spatial data. The key distinction is that spatial analysis has to consider spatial associations between artefacts.

Spatial data mining applications are decisive, for example, marketing, environmental analyses, risk mitigation, etc. For instance, a store may set its trading location, i.e., its clients' area, marketing, and evaluate their profile based on both its assets and the properties belonging to the area in which they reside [1].

Spatial knowledge mining is used for traffic risk analysis in our project. The probability calculation is focused on details on historical incidents coupled with thematic data on the road infrastructure, individuals, houses, etc. [2]. The project aims to define high-risk areas and examine and clarify these threats to the geographical neighbourhood. The infrastructure for spatial data mining explicitly facilitates these neighbourhood relations.

Geographic data analyses are currently focused primarily on standard statistics and multidimensional data processing and do not consider spatial data [3]. But the key characteristic of geographical data is that spatial observations appear to share identical (or correlated) attribute values. This forms the cornerstone of a different science field called "space statistics" which presupposes interdependence of neighbouring observations instead of conventional statistics [4]. There is abundant bibliography in this field, including popular geostatistics, recent advances in Anselin's Spatial Data Analysis (ESDA) and Openshaw's geographic analyzer (GAM). See Section 1.c of [5] for a description. To help contiguity, multidimensional empirical approaches have been extended [6]. As evidence-based assessments are presented, we maintain that spatial statistics is a part of spatial data mining. Any of them are now deployed in GIS or computational software [7].

Two big teams have contributed to the growth of spatial data processing data mining in databases. The first, a GeoMiner [8] extension, was established by the DB Research Lab (DB Research Lab) of the Simon Fraser University in Vancouver. The second (University of Munich) developed a neighborhood structure graph [9] based on certain algorithms. They also focused on the clustering framework based on hierarchical partitioning, grouping (extension of DBSCAN with an R*Tree), relationship laws (based on an effective spatial connection), characterization, and spatial patterns. They also developed the method of clustering. A hierarchical grid was used to optimize the clustering algorithm by STING (University of California) [10]. It is also possible to mention work on space data warehouses (University of Laval) [11].

This paper explains and illustrates the importance of data mining for geographical information structures in spatial data analytics performance. It will investigate both mathematical and database inference methods [12].

It has the following form. Section 2 describes and splits spatial data mining into generic functions. In section 3, spatial data mining approaches are defined in terms of these activities, whether extracted from the database field, statistics or artificial intelligence [13]. To underline their similarity and complementarity, we may equate the mathematical research method with the spatial database approach. Finally, we wind up learning about the subjects of science.
2. Definition of Spatial Data Mining

Space data mining (SDM) consists of retrieving information, space interactions and other resources not contained in the database directly. SDM is used for the detection of implicit regularities, spatial and/or non-spatial intercourse.

The feature of SDM is its spatial interaction. A spatial-temporal spectrum in which properties belonging to a specific location is usually related and clarified in respect of its vicinity's properties is a geographical index. In the research phase, we will also see the tremendous significance of spatial connections. Also, a key point is the temporal dimensions of spatial data, although they are hardly considered.

Data mining [14] are not spatial data suitable since they do not accept position data or tacit connections between objects. New approaches, including spatial relationships and spatial data care, must, therefore, be established. It takes time to approximate these spatial interactions, and the encoding of geometric position produces an immense amount of data. This uncertainty would impact global efficiency [15].

With GIS, you can query spatial data using programs or queries and carry out basic analytical jobs. GIS is not meant for complex data processing or the exploration of information [16]. They may not have general theoretical and rule-making approaches.

However, these current approaches need to be integrated and improved by integrating spatial data mining methods. For data access, spatial relations and presentation of graphic maps, GIS methods are essential. Only alphanumerical information can be produced by traditional data mining.

3. Spatial Data Mining Tasks

Space data mining activities typically expand data mining tasks to incorporate space data and parameters. These tasks include (i) data summarisation; (ii) grouping rules; (iii) clusters with related objects; (iv) data characterisation connections and dependencies to be found; and (v) anomalies since general patterns have been searched. They are conducted using various approaches, and others focused on mathematics and some in machine learning.

3.1. Spatial data summarisation

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

3.1.1. Statistical analysis of contiguous objects

3.1.1.1. Global autocorrelation. Elementary statistics, such as estimation of average, variation, etc. and visualisation methods, such as histograms and pie charts, are the most common way of summarising a data set. New methods for calculating worldwide neighbourhood dependence, such as local and local covariance, Geary's spatial auto-correlation, and Moran indexes, have been developed [17].

These methods are focused on the concept of an adjacent matrix which describes the inter-object spatial relationships. Note that this adjacency can fit various spatial relations, such as adjacency, distance difference and so forth.

3.1.1.2. Density analysis. This approach is included in the ESDA and does not include data information, contrary to the measure of autocorrelation. The idea is to measure the density in the space and visualize the point pattern by measuring each tiny circle window's intensity. The graphic approach could be defined.

3.1.1.3. Smooth, contrast and factorial analysis. Non-space features are overlooked in the density review. The study of geographic details typically concerns both alphanumerical and spatial features (called attributes). This involves incorporating spatial details throughout the research phase, with attributes and evaluating multi-attributes using multidimensional knowledge [18].

Two approaches alter attribute meaning from utilising the neighbouring matrix to integrating the spatial neighbourhood into attributes [19]. The first technique soothes each attribute value, substituting its neighbors' average value. The general aspects of the data are illustrated. The other compares data through deduction from each sum of this average.

Each attribute may then be evaluated using standard methods (called a variable) in statistics. However, multidimensional data processing approaches (e.g., factorial analyses) become important as several attributes (above tree) have to be evaluated jointly [20]. Their theory minimizes the number of variables by checking for the axes, where data values are maximally distributed. The association or dependencies between properties may be deduced by projecting and visualizing the axes' initial dataset.
The studied artefacts were initially considered autonomous in statistics and, in particular, in the above approaches. Several academic projects have been carried out [21] to look into space organisations. As the original table is converted using the smoothing or contradicting techniques, the application of factor analysis techniques to adjoining objects requires introducing traditional principal component analysis or correspondence analysis approaches.

3.1.2. Generalization. This mechanism raises the abstract degree of non-spatial features and reduces the geometric representation information by integrating neighbouring objects. It comes from the principle of induction based on attributes as defined in [22]. Here, the term hierarchy may involve spatial or non-spatial (thematic) hierarchy [23]. The following can be used as an example of a thematic hierarchy in agriculture: "cultivation style (food (maize, wheat, rice), plants, berries, etc.). This hierarchy may be applied directly by a field expert or created by an attribute-related deduction mechanism. There may also be a pre-existing spatial hierarchy, such as administrative boundaries, a geomechanical artificial split, such as a quad-trip [24], or a spatial clustering (see below).

Two forms are generalised: non-spatial, dominant generalisation, where first a thematic hierarchy is used, neighbouring items converge, and a dominant spatial generalisation, centred on spatial hierarchy first, preceded, with each widespread spatial meaning, by the accumulation or generalisation of non-spatial values. The algorithms' complexity in question is $O(N\log N)$, where $N$ is the current number of items.

3.1.3. Characteristic rules. The definition of the features typical of this specific section, but not of the entire database, has also been described in [25] as characterising its selected portion. Regarding a space database, the characteristics of objects and their neighbourhood properties to a certain degree are taken into consideration. Take a sub-set $S$ of empirical objects. The following parameters are used for this method: 1) value (relative frequency to the $S$ database); 2) trust (relation of artefacts in $S$ which reach the significance threshold in the neighbourhood).

This method throws up the properties $p_i = \langle\text{attribute, value}\rangle$, the relative frequency factors $freq-fac l$ (higher than the significance parameter) and the number $n_i$ of neighbours on which the property's frequency is extended. The following rule can express the characterisation:

$$S \not\ni p_1(n_1, freq-fac 1) \not\ni \ldots \not\ni p_k(n_k, freq-fac k).$$

3.2. Class identification

This job, also known as the controlled classification, offers a conceptual definition that generates the best database partitioning. The Laws of Classification are a decision tree where each node includes attribute criteria. The distinction between Spatial Databases is that this requirement may be a spatial predicate. The spatial artefacts rely on their neighbourhood; a rule could be applied to neighbourhood properties that contain the non-spatial properties of a specific entity.

The designation was used to study remotely sensed data primarily in spatial statistics to distinguish each pixel in a given group. To form a spatial object, homogeneous pixels would then be aggregated[26].

In spatial database approach [27] the description is used not only for the immediate neighbours, but also for the neighbour’s neighbours, and so on, as an arrangement of artefacts utilising both their properties (non spatial values) and the properties of their neighbours. Take the grouping of regions by their economic power as an example. The laws of classification are as follows:

High population $E$ neighbour = road $E$ neighborhood $= airport = >$ high economic power (95%)

GeoMiner may also be connected to a spatial attribute by classification criteria representing its inclusion in a region wider. The algorithm, whether through clustering or melting neighbouring objects, or a predefined spatial hierarchy could decide these areas.

This classification approach in GeoMiner is generalised by a new algorithm[28] into spatial predicates.

3.3. Clustering

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

3.3.1. Database approach. Paradoxically, relative to those in connection databases, clustering approaches for spatial databases do not seem quite innovative (automatic classification). The clustering takes place using an already classified similitudinal feature as semantical distance. Therefore, utilising the Euclidean distance to group adjacent items is natural in spatial datasets. Study work focused on algorithm optimisation. Geometric clustering establishes additional divisions such as residential placement of homes. This stage is mostly carried
out before other data mining activities, such as identifying a group's affiliation or characterisation within groups or other regional entities.

GeoMiner incorporates geometric clustering used in a point-set distribution and non-spatial generalisation. In the United States, we would want to characterise clusters of major cities and see if they are organised for starters. New areas corresponding to the convex hull of a community of cities would reflect the cluster's outcomes. Any points might remain beyond clusters, reflecting noise. For each defined attribute, a summary of each category can be created.

Many algorithms, like CLARANS, DBSCAN or STING, have been suggested for clustering. They also rely on optimising prices. GDBSCAN was illustrated in recent times in [29], a tool that applied more explicitly to spatial details. It refers not just to points of data to every spatial type and includes data attributes.

3.3.2. Statistic approach. This was specifically used in the epidemiological studies through clustering from an overview of points patterns [30]. This can be evaluated using a K-Function [4] in Openshaw's well-known Geographical Analysis Machine (GAM). The clusters can also be calculated by the two density estimates ratio: one of the sub-sets and one of the whole datasets.

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. The estimation of spatial dependency using the concept of spatial weight matrix [31] is concerned with the local self-relation. The discrepancy between the real spatial distribution of vector values and the random distribution will also be determined. It is also similar to a residual regression test. If the matrix is one column, a relation between one point and all the others is scanned.

3.4.2. extension to multi-level association rules. These comparisons may be widespread or comprehensive in the context of a defined hierarchy. Spatial hierarchies are refined or paired with a logical hierarchy of features (like the subdivision into regions, then departments and municipalities). The existing approach in connection bases includes grouping groups of subjects, to produce more general and specific laws, before searching for association rules. The fact that 64 per cent of homes are around 500 m from colleges, two-thirds are primary, and one-third secondary schools or high schools illustrates a hierarchy in a space-based database.

3.4.3. Group proximity rules. Suppose we have a community of private housing communities. The consumer wishes to specify that these category positions are described by those spatial artefacts closest to them. E.g., 65% of these homes are situated near lakes, beaches or mountains. It can be calculated.

The Partnership laws are proposed to be updated. This approach is focused on the exploration of entity classes that are also similar to predetermined types. An algorithm CRH is used to calculate an item's closeness to a group efficiently. Another algorithm named GenCom is then used to infer the laws of proximity and merge them with widespread attributes when a definition hierarchy 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 characterise spatial trends.

3.5.1. Database approach. The study will be carried out in four stages using the method mentioned in [32] based on central locations' principle. The former involves discovering centres by calculating a local maximum of particular features. The latter determines the theoretical trend by moving away from the centres. The latter determines the difference between these trends, and finally, we analyse the proprietary characteristics of those areas by explaining these trends. The pattern study of the unemployment rate relative to the distance from a city such as Munich is one example. The pattern analysis of house building growth is another example.

3.5.2. Geostatistical approach. Geostatistics is a space measurement technique and a predictor of Spatio-temporal phenomena. It was initially used for geological application (the geo prefix comes from geology). Geostatistics currently provides a class of technologies for studying and predicting unsafe values of variables spatially and/or temporally dispersed. These principles can be related to the climate. Structural research is the examination of such a connection. The "kriging" technique then predicts position values outside of the sample.

Bear in mind that geostatic is confined to the study of points or the polygonal subdivisions and discusses a particular variable or attribute. It represents a successful instrument for evaluating room and space-timing patterns, under certain circumstances.
4. **Comparison of SDM Approaches**

One of this thesis's purposes is that the entire academic body on the interpretation and retrieval of spatial data be taken together. The study was done either on the numbers or database learning, but they avoided one another much of the time. One must then be in a position to evaluate and examine them for the same empirical purpose. This segment would aim to compare all of these approaches and describe the points they have in common after classifying them by mission and differentiating among the different procedures that emerge from these deux approaches. This is a summary:

4.1. **Graphical methods and semantic methods**

Some approaches are focused exclusively on the graphical side of the results, such as in spatial analysis (density and relative cluster). Sometimes the effect is visual.

Others use a semantic description of spatial connections, including diagrams and surrounding matrices, on the other side. In addition to the clustering, which is a graphical process, the bulk of the database solution approaches come into this group. Auto-correlation checks, smoothing and smoothed factorial analysis as semantical approaches may be represented in the statistical approach.

4.2. **Taking account of contiguity**

In the usage of neighbourhood grammar, there are significant variations. Spatial relations are explicitly interpreted in the learning method as though they were a matter of assets. Conversely, these neighbourhood associations are either incorporated into models through a mathematical method, self-relation or employed as a smooth analysis to rectify the initial data.

These connections are also intra-thematic in the statistical approach, that is, among objects of the same theme. In contrast, in the learning approach, they can be inter-thematic (between different layers). This is particularly essential in an illustrating model, in which surrounding items, regardless of the subject, may interfere. For starters, rainfall and layers of population density correlate strongly.

Inter-theme connections of different spatial parameters are recovered using joint operators. Although these operators are time-consuming and complicated, they must be checked and optimized.

4.3. **Interpretation**

Furthermore, the research method, including generalisation, allows for the data to be combined with being summarised and synthesized. This method creates classifications with relatively little user interaction and provides relationship laws that are understandable by specialists.

Graphic approaches that are part of exploratory analyses are rather readable and need very little information.

Concerning factor analysis, it often synthesizes the data, but it does not minimise the number of items that can handicap vast quantities of data, compared to general analysis. The outcome will be of significant importance to an informed individual who can understand these methods but not to a neophyte in analyzing data.

4.4. **Complementarity**

These variations contribute to an incredibly important degree of complementarity from an empirical perspective. A generalisation process, for example, will allow the data to be condensed and streamlined to prepare for smooth or contrasting factor analyses. Until characterization, the quest for associations or grouping laws it will also be important to pursue generalisation. Similarly, to clarify localized concentration, characterization, or the quest of connections can be used.

The method mentioned in figure 1 is also different. To locate centers, the current pattern can be contrasted to a statistical trend to detect anomalies and then find properties that represent the locations of the deviations. It will require an evaluation of the density.
4.5 Accessibilities

Examples of easy calculation of public and private transit access by students are their distance from train stations, road distance and campus distance across the road network. Although almost 100 per cent of PR&H students lived less than 50 km away from a train station and a highway in 1999, our studies divided the station and highway proximity into different categories (e.g., 10 km and 5 km from stations roads). They analyzed the relevance of them for student registrations.

Thus cross-level classifications with multiple variables may be merged (for usability). Our survey found that most PR&H students in 1999 had easy access to public and private transportation (e.g., within 5 km to the station and 5 km to the highway). 1 - 2 3 - 5 6+ NSW Buffer 10km - 3 - 5 6+ VIC NSW Buffer 5km away from the station 5km Buffer 5km away (lower level of classification, 3b).

The distance from a specific campus (Thurgoona, Albury NSW) via road networks (shortest path) is lower than anticipated.

4.6 Socio-economic status

A student's home (the selected variables are average weekly income, and the standard of education) is typically special. More than 90% of the PR&H students in 1999 were from a region where the typical standard of schooling was rated as 'middle' - most citizens with either an affiliate or an undergraduate degree. The registration trend was also closely related to the average family weekly income. Near to 90% of PR&H students in 1999 came from the middle-income zone (with AUS$500 to AUS$999 weekly income). The study, as mentioned earlier's findings, may be interpreted as proof of the university marketing officers' prior experience.

4.7 Defining the Potential Markets

A collection of features can be established that are closely related to the PR&H student registrations of 1999. The future demand for 2000 may be found if all these characteristics are identified, including maps overlaid and areas intersected with these characteristics. This future market consists of gathering postcode zones, regional authorities or census regions.

5. Conclusion and Research Issues

Various spatial database data mining strategies were illustrated in this article, demonstrating two distinct academic communities: the statistics group, the database community and the research community have developed these tools. We also outlined and categorized this analysis and contrasted the two methods, stressing each method's specific effectiveness and future benefits to be integrated. This work marks the first move towards a methodology that incorporates the entire discovery process into spatial data bases and enables combining the methods as mentioned earlier as information mining. One approach is to consider spatial among other aspects while another is to consider how the linear or network form (such as roads) can affect specific graphical approaches. One approach is to analyse the temporality of spatial details. Anyway, the further development of the efficiency of these techniques remains important, which is the immense data quantities and another is the spatial interactions that are utilized intensively. These connections may be optimized using spatial indexes in the case of graphical approaches. Concerning the other approaches using neighborhoods, instancing the structure is expensive and as much as possible it is pre-computed.
References


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


