Geospatial Analysis Framework

Elisabeta Antonia Haller
"Mircea cel Bătrân" National College from Rm.Vâlcea
Carol I, 41, 240178, Râmnicu-Vâlcea, Romania
antonia_haller@yahoo.com

Abstract:

In a computerized society, the volume of data grows unexpectedly, making assessing their processing time a very difficult task. A priority has become the processing of data in useful information and knowledge. Thus we can say that data mining is a result of technological developments. The interpretation of spatial data has constituted a subject of research over time, reaching now a large variety of instruments and software products for representation and interpretation. What we need to understand beyond the facilities offered by one system or another, proprietary or open source solutions, is how they work and interact with spatial data.

Keywords: spatial data infrastructure, metadata, geospatial analysis, GIS Data Mining

1. Spatial data infrastructure

Like many other areas, over the last years, geographic fields recorded a significant increase in the collection of geographic data. There is a national and international program for the development of a spatial data infrastructure (SDI) that facilitates sharing and interoperability of geographical information. In this respect, the UE has adopted the INSPIRE Directive in May 2007, establishing an infrastructure for spatial information in Europe and which addresses 34 spatial data themes needed for environmental applications. If we say SDI, we say "metadata, spatial data sets and spatial data services; network services and technologies; agreements on sharing, access and use; coordination and monitoring mechanisms, processes and procedures, established, operated or made available in accordance with this Directive". [1]

With the SDI development programs the size of digital spatial data archives has increased, via the World Wide Web. In our days, geographic information are all over the Internet space: the virtual globe like NASA World Wind, released in mid-2004 or Google Earth, released in mid-2005, web map services like Google Maps from Google, OpenStreetMap from OGC, Bing Maps from Microsoft or Yahoo Maps and we mentioned just a small part of what we can find in the spatial data domain. Geographic data sharing has many inherent problems such as: handling different data formats, finding adequate data for use, differences in software versions, economical boundaries and often communication problems. There is *a standard for data exchange* meant to increase the efficiency in data distribution and sharing. We do not have standards only for geographic data, but also for metadata. We will here mention only a few data standards, the most known and used ones: *Spatial Data Transfer Standard* – SDTS, *Digital Geographic Information Exchange Standard* – DIGEST, *Topologically Integrated Geographic Encoding and Referencing* – TIGER.

Interoperability is one of the major concerns in GIS software. The connection between different GIS software and data transfer is made through standardized interfaces, where data is not converted or transferred. In order to transfer data from system S1 to system S2 we use a translator from system S1 to spatial data transfer standard and after the completion of the transfer we use another translator that converts data from standard format to S2 format. [5]

Open Geospatial Consortium (OGC), mainly, and national mapping agencies or international standard bodies make these standards possible. An important component of the geographic information infrastructure is spatial data clearinghouse. This is a distributed network of spatial data producers, managers and users electronically interconnected. Combining institutions with different software systems makes possible discovery, evaluation and downloading digital spatial data. The objective of a clearinghouse is to minimise effort for data capture and maximise the benefit of geographic information sharing.

Terms like geoportal, spatial data clearinghouse or SDI can not be mentioned without also mentioning *metadata*. This concept is used to describe both data and information, in terms of content and quality. With metadata, users find out data sets, purpose and scope of data in an easier way. “Data about data” should be flexible and ready to answer questions like who, what, when, where, why and how data are available. The answers are extremely important for users, metadata playing a variety of informative roles. These concepts (SDI, metadata and clearinghouse) together with others form the framework for the theory of geospatial analysis.

2. Geospatial analysis with data mining algorithms

When we say geospatial analysis we refer to a collection of techniques and tools for geographic analysis and GIS data processing software engines.

Images constitute extremely complex information and in its representation we take the decision, about which side will be involved, at what level of detail and in what period of time. There are two basic modes of representation: discrete representation of objects and representation with continuous fields. [7]

In discrete representation we take into account only well-defined objects in terms of shape, color and form. An important feature of discreet representation of objects is that they can be counted. In this sense a geographical area can be described as an area with landforms (mountains, plateaus, plains) or water surfaces (rivers, lakes, seas). The uncertainty arises when we have to differentiate a higher hill from a mountain or a water surface from a lake. In this way geographical objects acquire spatiality, dimensionality and area. They are called polygons and represent buildings, arable land. There are elements that have no area such as roads, railways, rivers and they are represented as lines, characterized by a single dimension. Moreover, there are zero-dimensional objects – individual, isolated elements. In fact, all these objects are perceived as three-dimensional objects by humans. In continuous representation we choose a finite number of measurable variables at any point on Earth's surface and we change values from one point to another. The continuous representation emphasizes elements that are difficult to fit into a table with attributes. Continuous fields are characterized by how they change and how abrupt is the change. In the computer representation of geographical data, the conceptual representation as discrete objects and continuous fields are strongly associated with vector data and raster data. The potential to perform spatial analysis and geographic knowledge discovery varies on the type of representation used.

According to Longley et all. [7, p320], the methods of spatial analysis can be divided in six categories. (Table 2.1).

Table 2.1 Types of spatial analysis

Spatial analysis method	Description
Queries	Retrieve information from database.
Measurements	Numerical value that describes geographic entities and relations between geographic entities.
Transformations	Changing, combining or comparing datasets.
Descriptive summaries	Descriptive statistics applied in GIS.
Optimization	*p-median problem* - selecting ideal locations according to well-defined rules
Hypothesis testing	Make generalizations about the whole from a sample dataset.

Spatial queries have different signification, depending on the suppliers. Information retrieval involves processes of selecting information by a query. Most of GIS applications facilitate this process of interrogation by a point-and-click visual interface. Users interact with database through spatial queries and the complexity of selection varies from an application to another. (Table 2.2)

Table 2.2 Spatial queries available on different platforms

Application	Description
ArcGis	Supports standard SELECT, JOIN and RELATE processes enabling tables to be joined and queried based on their relative locations, selects by location facility; does not support MAKE TABLE, UPDATE and DELETE
MapInfo	Supports standard SELECT and spatial join, without sharing a common field.
Manifold	Supports standard SELECT, specific spatial extensions (Adjacent, Contains, Intersection, etc), CREATE, UPDATE, DELETE; supports combinations with spatial operations (distance and buffer functions); supports combinations with operators that generate spatial objects (triangulation, convex hull of a point set); performs geometrical processing on the selected data.
PostGis	Supports standard SELECT, distance queries, spatial join, spatial indexes, coordinates re-projections inside the database

Measurements are made in all spatial databases. The most common queries include counting, area computations for a parcel, distance between two points and length of a river or road. If the model used for representing the world is the 2D Euclidean space, then the distance between two locations with Cartesian coordinates (x_1, y_1) and (x_2, y_2) is calculated through formula (1). The length of a polyline is the sum of the constitutive segments, where each length of a segment is calculated through (1).

$$d_{12} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \; (1)$$

If the model used for representing the world is in spherical space, the distance between points 1 and 2 is calculated with the Cosine formula (2) or with a more accurate one, the Haversine formula (3).

$$d_{12} = R\cos^{-1}[\sin\varphi_1 \sin\varphi_2 + \cos\varphi_1 \cos\varphi_2 \cos(\lambda_1 - \lambda_2)] \; (2)$$

$$d_{12} = 2R\sin^{-1}\left(\sqrt{\sin^2\left(\frac{\varphi_1 - \varphi_2}{2}\right) + \sin^2\left(\frac{\lambda_1 - \lambda_2}{2}\right)\cos\varphi_1 \cos\varphi_2}\right) \; (3)$$

where R is the Earth's radius and (φ_1, λ_1) is the latitude and longitude for the first point, and (φ_2, λ_2) is the latitude and longitude for the second point.

Another type of measure for GIS entities is their shape and it is calculated through (4).

$$S = \frac{P}{2\sqrt{\pi A}} \; (4)$$

where P is the perimeter length and A is the area.

On the raster data measures are obvious because the cells are regular. The only thing that could vary is the cell (pixel) resolution. A line is defined like a cells collection with common features.

The *transformations* process combines spatial data layers and produces a new one. Operations that fall under the incidence of transformations are: buffering (vector, raster, hybrid or network), point in polygon, polygon overlay and spatial interpolation,

Descriptive summaries try to find patterns, geographical distribution and phenomena using simple statistic methods. The bases of building descriptive summaries are centroids, dispersion, labeled/unlabeled features and fractional dimension.

The role of *optimization* is to create improved design and find out answers for questions like point location, routing problem and optimum paths.

A *hypothesis test on geographic data is* a particular case of statistical inference. Statistical inference is based on the theory that we can generalize from a small group to a large context. Geographic data are natural experiments and we can not randomly extract samples from this data because they are spatially dependent in the area of interest. "*Spatial analysis is often conducted on all of the available data, so there is no concept of a universe from which the data were drawn, and about which inferences are to be made.*" [2]

Geographic information could be more valuable when we use neural networks, fuzzy logic or others AI tools for extracting and evaluating data. The first major step for GIS was the development of spatial databases. After this important step, discovering useful information with one important artificial intelligence tool was almost a necesity because of the large size of databases with geographic data: spatial data mining. Data mining techniques are useful in all types of interpretations of geographic data: land use, image analysis, pattern analysis, wheather and floods prediction, forestry, oceanography, transportation, bio-sphere studies or environment changes. In almost every case where GIS is being used, AI tools could enchancing decision-making process [9].

A relevant part of data mining techniques that is used for geographic knowledge discovery is spatial clustering. Clustering is the process of grouping physical or abstract data into classes of similar objects. In a cluster we find data objects that are similar to each other and dissimilar with the objects in other clusters. In clustering we do not need to label large set of patterns [2].

Spatial clustering could be used in pattern recognition, data analysis and image processing. Identification of areas of similar land use in an earth observation database, finding groups of different objects in an area according to geographical location could be only a few examples of spatial clustering.

There are three main divisions for clustering: partitional clustering, hierarchical clustering and locality-based clustering. In all this types of clustering we can found algorithms which can be successfully applied to geographic data (Figure 1). [8]

Geographic data can be found in different forms. The most used data is *point* data and can represent buildings, roads centers or lands. Moreover, data objects are not static and in present we can keep track of moving objects like vehicle position data or animal movement data. For this type of data, called *trajectory* data, machine learning field has developed powerful algorithms. A *trajectory* is a set of multidimensional points. In this context, a *cluster* is a set of *trajectory partitions*. A trajectory partition is a line segment $p_i p_j, i < j$, where p_i and p_j are points from the same trajectory. Finally, a *representative trajectory* is an imaginary trajectory that indicates the behavior of the line segments in a cluster [2].

We must notice that early clustering methods do not use trajectory partitioning. J-G Lee, J. Han and K-Y Whang has proposed a partition-and-group framework for trajectory clustering in order to resolve the problem of discovering clusters of sub-trajectories [10]. They proposed a density-based method, a partial trajectory clustering algorithm named TRACLUS (TRAjectory CLUStering). If we look at the trajectory clusters as a whole is possible to miss common behavior.

The solution proposed in [10] is to partition trajectories into line segments and then group similar line segments. Trajectory clustering based on partition-and-group framework consists of two phases:

- partitioning phase: trajectories are partitioned into line segments
- grouping phase: grouping similar line segments into a cluster

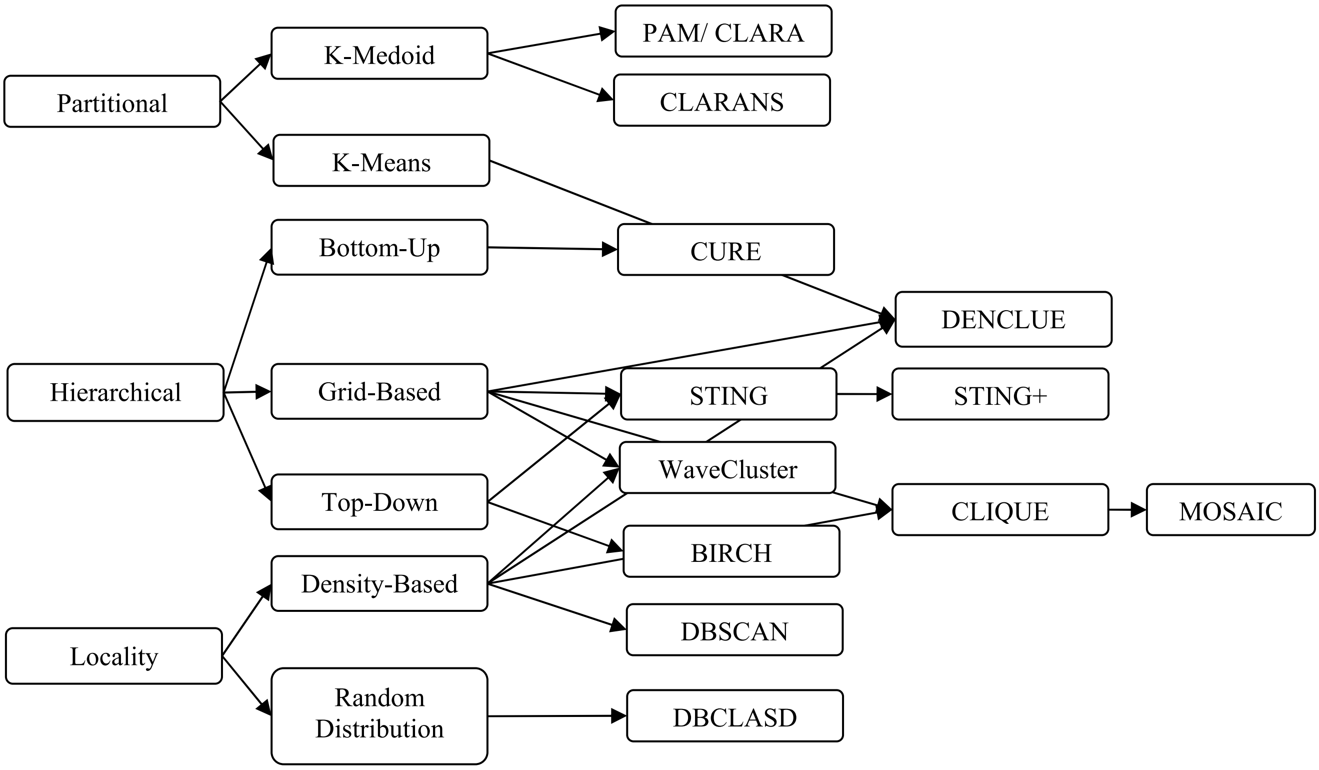

Figure 1. Different types of spatial clustering algorithms

The TRACLUS algorithm is below:

```
Algorithm TRACLUS (TRAjectory CLUStering)
INPUT: $\mathrm{T} = \{TR_1, TR_2, \ldots, TR_{num_{tra}}\}$, a set of trajectories;
OUTPUT: (1) $\mathrm{O} = \{C_1, C_2, \ldots, C_{num_{clus}}\}$, a set of clusters;
         (2) A set of representative trajectories;
ALGORITHM:
//Partitioning phase
For each $(TR \in \mathrm{T})$do
      Execute Approximate Trajectory Partitioning;
      Get a set $L$ of line segments;
      Accumulate $L$ into a set $D$;
//Grouping phase
Execute Line Segment Clustering for $D$;
Get a set $O$ of clusters as the result;
For each $(C \in O)$ do
      Execute Representative Trajectory Generation;
      Generate representative trajectory as the result;
```

This algorithm can solve a wide range of problems that involves trajectory data like: common behaviors of hurricanes that improve the accuracy of forecasts, vehicle traffic management, supply chain management, effects of roads and traffic in animal movement or military and security markets.

3. Conclusion

The great advantage of the latest research in geospatial analysis is the use of data mining techniques. Finding patterns, anomalies and clusters in spatial data is a challenge for both geographers and mathematicians. The new trend for opening in the WWW space makes exploring data possible at other levels. The GIS software provides the opportunities for exploring the environment and develops new analytical solutions to geospatial problems.

4. References

[1] INSPIRE official site, from http://inspire.jrc.ec.europa.eu/index.cfm.

[2] Miller, J. H., Han, J. (2009). *Geographic Data Mining and Knowledge Discovery*, Second Edition, CRC Press.

[3] Smith, Goodchild P. A., Longley, P.A. (2009), *Geospatial Analysis - a comprehensive guide*, 3rd edition, [Electronic version] Retrieved from http://www.spatialanalysisonline.com.

[4] Guo, D., Gahegan, M., McEachren, M. A. (2003) *An Integrated Environment for High-dimensional Geographic Data Minig*. GeoVista Center.

[5] de Bay, A., Knippers, R.A., Yuxian, S. (2000). Principles of Geographic Information System. The International Institute for Aerospace Survey and Earth Science.

[6] Verbila, D. L. (2003) *Practical GIS Analysis*. Taylor and Francis.

[7] Longley, P.A., Goodchild, P.A., Maguire, D.J., Rhind, D.W (2005). *Geographical Information System and Science*, 2nd edition. John Wiley and Sons Ltd.

[8] Kolatch, E. (2001). Clustering Algorithms for Spatial Databases: A Survey. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.1145.

[9] Thurston, J. (2002). GIS & Artificial Neural Networks: Does Your GIS Think? Retrieved from http://integralgis.com/pdf/Neural%20Networks.pdf

[10] Lee, J-G., Han, J., Whang K-Y (2007). Trajectory clustering: A partition-and-group framework. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.76.8098.