An approach for Mining Complex Spatial Dataset

Grace L. Samson[1], Joan Lu[1], Lizhen Wang[2], Dave Wilson[1]

[1]Informatics, School of Computing and Engineering, University of Huddersfield; Huddersfield, UK
[2] School of Information Science and Engineering, Yunnan University, China PRC

Abstract: *Spatial data mining organizes by location what is interesting as such, specific features of spatial data mining (including observations that are not independent and spatial autocorrelation among the features) that preclude the use of general purpose data mining algorithms poses a serious challenge in the task of mining meaningful patterns from spatial systems. This creates the complexity that characterises complex spatial systems. Thus, the major challenge for a spatial data miner in trying to build a general complex spatial model would be; to be able to integrate the elements of these complex systems in a way that is optimally effective in any particular case. We have examined ways of creating explicit spatial model that represents an application of mining techniques capable of analysing data from a complex spatial system and then producing information that would be useful in various disciplines where spatial data form the basis of general interest.*

Keywords: Spatial data; Complex systems; Patterns mining; Spatial models; Spatial database

1 Introduction

Spatial data mining is the quantitative study of phenomena that is located in space. This means that there is an explicit consideration of the location and spatial arrangement of the object to be analysed [9].We have focused on the unique features that distinguish spatial data mining from classical data Mining, and present major accomplishments of spatial data mining research, especially regarding predictive modelling, spatial outlier detection, spatial co-location rule mining, and spatial clustering. Spatial data mining organizes by location what is interesting therefore, the main purpose of spatial data mining is to search for interesting, valuable, and unexpected spatial patterns; which can be useful in so many application domains. Most often than not the pattern discovered always provide a new understanding of the real world as such, the search must be a non-trivial one and should be as automated as possible with a large search space of plausible hypothesis. Attention to location, spatial interaction, spatial structure and spatial processes lies at the heart of activities in several disciplines today and as such demands the urgent development of tools capable of analysing and managing such data which typically can only be represented by means of geometric features, for instance, consider the examples of spatial data described by [23] as (a) Percentage cover of woody plants along a line division; (b) Land cover from some rangeland types within a specified area of a coastal region; these include some special cases of spatial data. Finding implicit regularities, rules or patterns hidden in spatial databases is an important task for example in geo-marketing, traffic control or environmental studies [6]. The ultimate goal of spatial data mining is to integrate and further extend methods of traditional data mining in various fields for the analysis and management of large and complex spatial data. The underlying concept is based on the fact that spatial data types (e.g *points, lines, polygons and regions*) are not supported by the *conventional database management system*. Studying spatial data management helps us to discover the relationship between spatial and non-spatial data and to be able to build and query a spatial knowledgebase.

1.1 Related research

Geospatial data is the data or information that identifies the geographic location of features and boundaries on earth (such as natural or constructed features), oceans etc. Spatial data are usually stored as *co-ordinates* and *topology* that can be mapped. They are often accessed, manipulated and analysed through geographic information system. Spatial data mining and geographic knowledge discovery has emerged as an active research area focusing on the development of theory, methodology, and practice for the extraction of useful information and knowledge from massive and complex spatial databases, Therefore, there is an urgent need for effective and efficient methods to extract unknown and unexpected information from spatial data sets of unprecedentedly large size, high dimensionality, and complexity [18]. According to [13] geographic information systems contain high level spatial operators that are uncommon in conventional database management system (DMS). This has led to an increased development of research issues that focus on technologies, techniques and trends that identifies properties that a spatial data model, dedicated to support spatial data for cartography, topography, cadastral and relevant applications, should satisfy. These properties concern the data types, data structures and spatial operations of the model [22]. In their work, [15] asserted that for every spatial data object, the attribute data are referenced to a specific location; which means that they are highly dependent on location and also influenced by neighbouring object (which has given rise to the mining of collocation pattern between spatial objects). Existing DBMS do not support complex spatial relations that exist between spatial objects thus to achieve this, the functionalities of the DBMS should be extended to incorporate the facilities of these complex spatial relations into their query language by providing for the DBMS a model of how to process and optimize queries over spatial relations [4]. Spatial database management refers to the extraction of implicit knowledge, spatial relations or other patterns not explicitly stored in spatial databases. Traditional data organisation and retrieval tools can only handle the storage and retrieval of explicitly stored data [15]. The importance of handling a spatial database derives from the need to deal with geometric, geographic or spatial data (i.e data related to space). According to [22] one remarkable feature of a spatial database is based on the fact that the management of geographic data is split into two distinct types of processing, one for the spatial data and another for the attributes of conventional data and their association with spatial data. In other words, according to [12], spatial database systems deals with the fundamental database technology for geographic information systems and other applications and querying this database is to connect the operations of a spatial algebra (including predicates to express spatial relationships) to the facilities of a DBMS query language.

2 Methods of mining spatial patterns

The major activities involved in mining spatial patterns include:

1. Dataset Preparation

2. Initial Data Exploration

3. Predicting Process And Mining Predictions

In this work, we looked at modelling (PREDICTIVE), querying and implementing a spatial database for event prediction using basic spatial data mining algorithm.

The term spatial database system is associated with a view of a database as containing sets of objects in space rather than images or pictures of a space. Consequently, mining spatial patterns from a complex spatial system would basically involve the description of the two categories of data obtainable in all geographic data (i.e spatial data and attribute data). In doing this, some of the major issues to consider include: data description, data manipulation and data representation.

2.1 Data description

2.1.1 Describing spatial data

- Properties of location in a map are often *"autocorrelated" (patterns exist)*

- Spatial data types are *complex (e.g points, lines and polygons)*

  Spatial data *denotes continuous feature*

- *Spatial operators include (overlay, re-class, distance etc.)*

2.1.2 Describing non-spatial data

- Data deals with simple domains e.g *numbers and symbols*

- Data *describe discrete object*

  Data *are independent* of each other

These descriptions identify properties that a spatial data model, dedicated to support spatial data for cartography, topography, cadastral and relevant applications, should satisfy. These properties concern the *data types, data structures and spatial operations* of the model as listed below:

- *Spatial operations (spatial query, layering/overlaying, buffering)*

- *Spatial data* which describes location (where)

- *Attribute data* which specifies characteristics at that location (what, how much, and when)

2.2 Spatial Data Representation

Representing spatial data in the form that the computer would understand requires grouping the data into layers according to the individual components with similar features (example layer could be waterlines, elevation, temperature, topography e.t.c).

In general, two distinct data structures are considered when representing spatial data digitally these include; (i) raster data structure (ii) vector data structure.

2.2.1 Raster data structure

According to [21], raster cell is usually a square, but could theoretically be another regular polygon that is able to fully cover an image area without leaving holes in the covered region, e.g. a triangle, hexagon or rectangle. Raster data structure according to [11], is similar to placing a regular grid over a study region and representing the geographical feature found in each grid cell numerically: for example, 1 for loamy, 2 for clay and so on. A raster consists of a matrix of cells (or pixels) organized into rows and columns (or a grid) where each cell contains a value representing information, such as temperature (as you can see in the figure below)

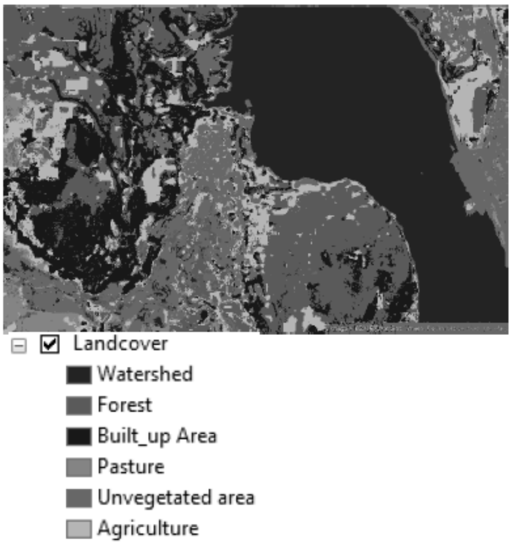

Figure 1: example of a raster data representation

2.2.1 Vector data structure

Vector data structure represents geographic objects with the basic elements *points*, *lines* and *areas*, also called polygons. From the description given by [11], vector data is based on recording point locations (zero dimensions) using *x and y coordinates, stored within two columns of a database.* By assigning each feature a unique ID, a relational database can be used to link location to an attribute table describing what is found there.

2.3 Spatial Data Types

[25] Observed that in trying to discover pattern in real world data, the different models in which real world data is organised and the pattern discovery technique to be applied to this models must be considered. Data types of a spatial set are the major element of a spatial database as we have described by the examples below.

Continuous data types: elevation, rainfall, ocean salinity

Areas data types:

- *unbounded:* land-use, market areas, soils, rock type
- *bounded:* city/county/state boundaries, ownership parcels, zoning
- *Moving:* air masses, animal herds, schools of fish

Networks data types: roads, transmission lines, streams

Points data types:

- *fixed:* wells, street lamps, addresses
- *moving:* cars, fish, deer

2.4 Data manipulation

The main application driving research in spatial database systems are GIS. Hence we consider some modelling needs in this area which are typical also for other applications. Examples are given for two dimensional *space (length and breadth)*, but almost everywhere, extension to the three - or more-dimensional case is possible. There are two important alternative views of what needs to be represented:

- *Objects in space*: in this case, we are interested in distinct entities arranged in space each of which has its own geometric description allows us to model, for example, *cities, forests, or rivers*

- *Space*: here, we wish to describe space itself that is describing every point in space. Models thematic maps describing e.g. *land use/cover* or the partition of a *country into districts.*

3 Knowledge discovery task in spatial data mining

The essence of data mining is to demonstrate the possible contribution of general KDD methods that are not specifically designed for spatially referenced data. Knowledge discovery in a *spatial database* involves finding *implicit regularities*, *rules* or *patterns* hidden in spatial databases. These are grouped under several basic categories in terms of the kind of knowledge to be discovered. Spatial data mining encompasses various tasks and, for each task, a number of different methods are often available, which could be *computational*, *statistical*, *visual*, or some combination of them. Some common spatial data mining task includes:

- *Spatial classification/prediction*
- *Spatial association rule mining*
- *Spatial cluster analysis*
- *Geo-visualization e.t.c*

These tasks can generally be classified into two categories:

- Modelling and
- Querying of a given spatial database

In general, the major tasks a spatial data miner may face would include of these (as shown in table 1):

Table 1: Example tasks in spatial pattern mining

Location Prediction
Predict that is trying to identify where a phenomenon will occur.

- predicting location of protein sub cellular [5]
- Predicting location of a mobile cellular networks user [1]

Spatial Interactions
The researcher is trying to find out which subsets of spatial phenomena interact?

- Application of spatial information to mobile computing [8]
- Applying spatial interactions to the analysis of crime incidents [14]

Hot spot -Finding which locations are unusual or share commonalities through spatial clustering

- Detecting spatial hot spots in landscape ecology [19]
- *Spatial* Organization of DNA in the Nucleus May Determine Positions of Recombination Hot Spots [25]
- Applying clustering techniques to crime hot-spot analysis [7]
- Other application areas include earthquake analysis, vehicle crashes, agricultural situations …..

Spatial outliers' detection
Trying to identify abnormal patterns (outliers) from large data sets

- *Detecting Outliers* in Gamma Distribution [20]
- *Bearing Based Selection* in Mobile Spatial Interaction [27]

4 Results and Discussion

[2] Has established that the data inputs of spatial data mining are more complex than the inputs of classical data mining because they include extended objects such as points, lines, and polygons. The data inputs of spatial data mining have two distinct types of attributes: non-spatial attribute and spatial attribute. Non-spatial attributes are used to characterize non-spatial features of objects, such as name, population, and unemployment rate for a city. They are the same as the attributes used in the data inputs of classical Data Mining. Spatial attributes are used to define the spatial location and extent of spatial objects The spatial attributes of a spatial object most often include information related to spatial locations, e.g., longitude, latitude and elevation, as well as shape.

One feasible way to deal with implicit spatial relationships is to materialize the relationships into traditional data input columns and then apply classical data mining techniques - although the materialization may result in loss of information.

However, in [26], it was also established that spatial context such as *autocorrelation* is the key challenge in spatial data mining especially in the area of spatial classification. And then we saw the most obvious challenge of spatial data mining (which is a general problem in field on data mining) in [28] as missing data. [28] acknowledged that since data mining process deals greatly with the development of association rule, patter recognition, classification, estimation and prediction, it will be very pertinent to have serious concern on the accuracy of the database to be modelled and on the sample data chosen for building a training set, in other words, the issue of *missing data* must be addressed since ignoring this problem can lead to a partial judgement of the models being evaluated and then finally lead to inaccurate data mining conclusions.

4.1 Data selection

- Measuring per cent occurrence of objects from digital images can save time and expense relative to conventional field measurements [3]

- Ecological assessments incorporating ground-cover measurements (as shown in figure 2) have relied on point sampling using point frames [16]; [17]

4.2 Data preparation

In addition to the DM process, which actually extracts knowledge from data, KDD process includes several pre-processing (*data preparation*) and post-processing (*knowledge refinement*) phases [10].

4.3 Points sampling

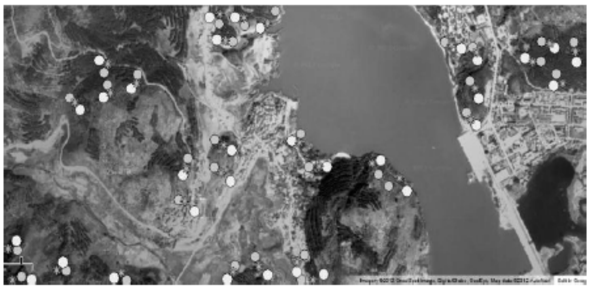

Figure 2: showing locations where sample point where selected on our base-map

4.3.1 Sampling methods

"The main aim of the analysis of mapped point data is to detect patterns (i.e., to draw inference regarding the distribution of an observed set of locations)" [29]. We have adopted the *sampling* method of data collection, because we are dealing with data that change across a surface over a period of time e.g temperature, precipitation, and so on. According to [3], measuring per cent occurrences of objects from digital images can save time and expense relative to conventional field measurements. Also, [30] established that ecological assessments incorporating ground-cover (the area, usually expressed as a percentage, of ground covered by the vertical projection of vegetation, litter, and rock) measurements have relied transect methods.
The measurement of ground cover from images has several potential advantages, including acceleration of field work, increased flexibility, repeatability, and convenience in the time and place actual measurements are made.

5 Conclusion

Spatial data mining is a branch of data mining where space and location of object is an important factor. In work, we have carried out an extensive research on the field of data mining and we have developed a framework for spatial data mining which is suitable for further expansion and research. We looked at the various branches and tools for data mining and we had a detailed study of spatial data mining; tools techniques, methods, and tasks. We also looked at the various application areas of spatial data mining and the nature of specific pattern that could exist in a given spatial dataset.

References:

[1]. Anagnostopoulos, T., Anagnostopoulos, C. and Hadjiefthymiades, S. (2012) "Efficient Location Prediction in Mobile Cellular Networks",

International Journal of Wireless Information Networks. 19 (2), pp. 97-111.

[2]. Bolstad, P. (2002). GIS Foundamentals: A Fisrt Text on GIS. Eider Press.

[3]. Booth, D. T., Cox S. E., and Berryman, R. D. (2006) "Point Sampling Digital Imagery with 'Samplepoint'" Environmental Monitoring and Assessment Springer. 123, pp. 97–108

[4]. Clementini, E., Sharma, J., and Egenhofer M. J. (1994) "Modelling Topological Spatial Relations: Strategies for query processing" Computer and graphics. 18 (6), pp.815 – 822.

[5]. Chou, K. and Shen, H. (2007) "Recent progress in protein subcellular location prediction", Analytical Biochemistry, 370 (1), pp. 1-16

[6]. Esther, M., Kriegel, H. and Sander, J. (2001) "Algorithm and Application for Spatial Data Miming" Geographic data Information and Knowledge Discovery Research Monograph in GIS.

[7]. Estivill-Castro, V. and Lee, I. (2002) "Multi-level clustering and its visualization for exploratory spatial analysis" GeoInformatica, 6 (2002), pp. 123–152

[8]. Fröhlich, P., Simon, R., Baillie, L., Roberts, J. and Murray-Smith, R. (2007) "Mobile spatial interaction", ACM. pp. 2841

[9]. Gatrell, A.C. and Bailey, T.C. (1995) Interactive spatial data analysis, Harlow: .Longman Scientific &Technical.

[10]. Ghosh, A. and Freitas, A., A. (2003) "Guest editorial data mining and knowledge discovery with evolutionary algorithms", IEEE Transactions On Evolutionary Computation. 7 (6), pp. 517 - 518.

[11]. Gregory, D., Johnston, R. and Pratt, G. Eds. (2009) Dictionary of Human Geography. 5th ed. Hoboken, NJ, USA: Wiley-Blackwell. [Online] Available at :http://site.ebrary.com/lib/uoh/Doc?id=10308208&ppg=816 [Accessed 18 August 2012]

[12]. Güting, R., H. (1994) "An introduction to spatial database systems" The International Journal on Very Large Data Bases. 3 (4), pp. 357 – 399

[13]. Gunther, O. and Buchmann, A. (1990) "Research Issues in Spatial Database" SIGMOD RECORD. 19 (4), pp.61-68

[14]. Kakamu, K., Polasek, W. and Wago, H. (2008) "Spatial interaction of crime incidents in Japan", Mathematics and Computers in Simulation. 78 (2), pp. 276-282.

[15]. Koperski, K. and Han, J. (1995) Discovery of spatial association rules in geographic information databases. In Proceeding of the 4th Int'l Symposium on Large Spatial Databases (SSD'95): Portland, Maine, Aug. pp 47-66.

[16]. Levy, E. B. (1927) "Grasslands of New Zealand", New Zealand Journal of Agriculture 34, 143–164.

[17]. Levy, E. B. and Madden, E. A. (1933) "The point method of pasture analysis", New Zealand Journal of Agriculture. 46, pp. 267–269.

[18]. Mennis, J. and Guo, D. (2009) "Spatial data mining and geographic knowledge discovery—An introduction",Computers, Environment and Urban Systems. 33 (6), pp. 403-408

[19]. Nelson, T.A. and Boots, B. (2008) "Detecting Spatial Hot Spots in Landscape Ecology", Ecography, 31 (5), pp. 556-566

[20]. Nooghabi, M. J, Nooghabi, H. J. and Nasiri, P. (2010) "**Detecting Outlier**s in Gamma Distribution" Communications in Statistics - Theory and Methods. 39 (4), pp. 698 - 706

[21]. Neuman, A., Freimark, H. and Wehrle, A. (2010) "Geodata Structures and Data Models" [online] Available at: https://geodata.ethz.ch/geovite/ -Version September 2010. [Accessed 12th August 2012]

[22]. Papadopoulos, A.N., Manolopoulos, Y. and Vassilakopoulos, M.G. (2004) Spatial databases: technologies, techniques and trend. US: Idea Group

[23]. Perry, J. N., Liebhold, A. M., Rosenberg, M. S., Dungan, J., Miriti, M., Jakomulska A., and Citron-Pousty S. (2002). "Illustrations and guidelines for selecting statistical methods for quantifying spatial pattern in ecological data" ECOGRAPHY. 25, pp. 578–600

[24]. Pudi, R. and Krishna, P. R. (2009) *Data Mining*. India: Oxford University Press

[25]. Razin, S.V. and Larovaia, O.V. (2005) "Spatial Organization of DNA in the Nucleus May Determine Positions of Recombination Hot Spots", Molecular Biology. 39 (4), pp. 543-548.

[26]. Shekhar, S., Schrater, P.R., Vatsavai, R.R., Weili Wu and Chawla, S. (2002) "Spatial contextual classification and prediction models for mining geospatial data". pp. 174.

[27]. Strachan, S. and Murray-Smith, R. (2009) "Bearing-based selection in mobile spatial interaction", Personal and Ubiquitous Computing. 13 (4), pp. 265-280.

[28]. Wang, J. eds. (2003) Data Mining: Opportunities and Challenges. US: IGI Global

[29]. Waller, L. A. and Gotway, C. A. (2004) *Applied Spatial Statistics for Public Health Data.* New York: Wiley

[30]. Interagency Technical Team (ITT): 1996, Sampling Vegetation Attributes, Interagency Technical Reference, Report No. BLM/RS/ST-96/002+1730. Denver, CO: U.S. Department of the Interior, Bureau of Land Management – National Applied Resources Science Centre. [Online] Available at:http://www.blm.gov/nstc/library/pdf/samplveg.pdf. [Accessed 22 Sept. 2012]