The Study of Knowledge Discovery with Spatial Data Mining in Epidemiology Database

Siddhi Nath Rajan[1], Ashok K Sinha[2], J B Singh[3]

[1] IMS Engineering College, MTU, Noida, INDIA, Research Scholar Shobhit University, Meerut,
[2] Prof & Head of Dept. (IT), ABES Engineering College,Ghaziabad,INDIA, External Supervisor,
[3] Professor and Internal research Supervisor, Shobhit University, Meerut, INDIA,

Abstract

Large amount of spatial data is being collected in various applications like remote sensing, computer cartography, geographical information system (GIS), environmental assessment and planning, etc. Now the real challenge is to discover interesting, implicit, and previously unknown knowledge from this large database. This is what the objective of spatial data mining. The real work is to extend the scope of data mining from relational and transactional database to spatial database and apply it in the study of spatial distribution of epidemiology. The paper summarizes the work that has been done so far in spatial data mining from spatial data generalization, mining spatial association rule to spatial data clustering.

Key Words: KDD, Spatial Data Type, Spatial Data Mining, Raster Map, Vector Map.

1. Introduction

The spatial data mining in spatial distribution of field specific database is an interdisciplinary research area which basically focuses on knowledge discovery from heterogeneous format of spatial database. The study focuses *first* on the different format of spatial data mining techniques and *secondly* a suitable technique to work on spatial distribution of epidemiology database to mine knowledge from such database.

The rapidly growing data creates the necessity of knowledge / information discovery from data which leads to promising emerging field, called the data mining or knowledge discovery from database (KDD). Spatial Data Mining, Shekhar & Chawla 2003[1], describes as a process of discovering previously unknown, but potentially useful patterns from spatial database. The process of data mining could be the integration of many things including machine learning, database system, statistics, and information theory. There are many studies available of data mining in relational and transactional database [2,3,4,5], the concept is in high demand to apply it in many other applicative area like spatial database, temporal database, multimedia database, object-oriented database etc. Section 2 discusses various methods and research gap of discovering interesting knowledge from spatial data whereas section 3 discusses one of the applicative are such as applying spatial data mining in spatial epidemiology database. Section 4 discusses the future direction of the research work.

2. Spatial data mining

Spatial data are the data related to objects that occupy space. It contains topological and/or distance information and is often organized by spatial indexing structures and accessed by spatial access methods. The objects stored in spatial database are the spatial objects represented by spatial data type and are having implicit relationship among them. The implicit relationship among the objects and the distinct feature of spatial database poses challenge and bring opportunities for mining information from spatial data [6]. *Knowledge discovery from database refers to the extraction of implicit knowledge, spatial relation, or other patterns not explicitly stored in spatial database*[7].

The work related to statistics[8,9,10,11], machine learning[12,13,14] and database systems[15,16] laid the foundation of knowledge discovery from database. Then after, with respect to spatial database, the study related to computational geometry[5],spatial data structure[17,18,19] and spatial reasoning [20,21] paved the way for the study of spatial data mining.

The statistical spatial analysis [9,11] has been the most common approach for analyzing spatial data. It handles very efficiently the numerical data which comes from the realistic model of spatial phenomena. But the assumption of statistical independence among the spatial distributed data causes problem as many of the spatial data are in fact interrelated. It is because the spatial objects are influenced by their neighboring objects. At the same time the statistical approach cannot model non linear rules very well. Statistical methods also do not work well with incomplete or inconclusive data. Another problem related to statistical spatial analysis is the expensive computation of the result. To supplement the work the machine learning techniques [12,14] and the spatial database potential[22,23] was nicely utilized. Now to model the non linear rules out of the spatial and non spatial data the potential of soft computing can be used.

2.1 Spatial Data Mining[SDM] Components

Following are some important attributes in the study of spatial data mining:

Rules: With the help of combined approach of SDM techniques, various rules can be discovered such as spatial association rule, spatial characteristic rule, deviation and evolution rule, and discriminate rule etc.

Thematic Map : It is a map that shows a theme, which is a single spatial distribution or a pattern, using a specific map type[2]. It presents the spatial distribution of a single or a few attributes. Spatial classification is one of the techniques that analyze spatial and non-spatial attributes of the data objects to partition the data into a set of classes. These classes generate a map representing groups of related data objects. There are two ways to represent thematic maps: *raster map* and *vector map*. The raster image thematic maps

have pixels associated with the attribute values. In the vector form the spatial objects are represented by its geometry i.e. boundary representation and thematic attributes.

Image Databases: These are special kind of spatial databases which consists of images and pictures. They are stored in the form of grid array representing the image intensity in one or more spectral ranges.

2.2 SDM architecture, spatial data structure

There are various architectures proposed for data mining. Some of the important architectures are J Han & Y. Fu's[24] architecture DBLEARN/DBMINER, M. Holsheimer and M. Kersten's [25] parallel architecture and C. J Mathu & Chan's[26] multi component architecture. These are the general data mining prototypes but they can be used or extended to handle spatial data mining. Mathu's architecture is very general and has been used by other researchers in spatial data mining, including M. Ester etc. [27]. The important spatial operations like spatial joins, map overlays, nearest neighbor queries are some important spatial operators. Thus in order to work efficiently, the operators requires an efficient spatial access method (SAM) and appropriate spatial data structure. Spatial data structure consists of points, lines, rectangles etc. and to build indices for these data , multidimensional trees have been proposed such as quad tree[46], k-d tree, R-trees, R* -tree etc.

2.3 Knowledge Discovery Methods.

The spatial database consists of spatial objects and non-spatial description. The non-spatial description of the spatial object can be stored in the traditional relational database[22]. There are two different properties of spatial data and they are geometric and topological.. The geometric properties could be spatial location, area, perimeter etc. whereas topological properties can be adjacencies, inclusion etc. The figure below (Figure: 1) describes how the non-spatial and spatial attribute values about the states of India are mapped in a database.

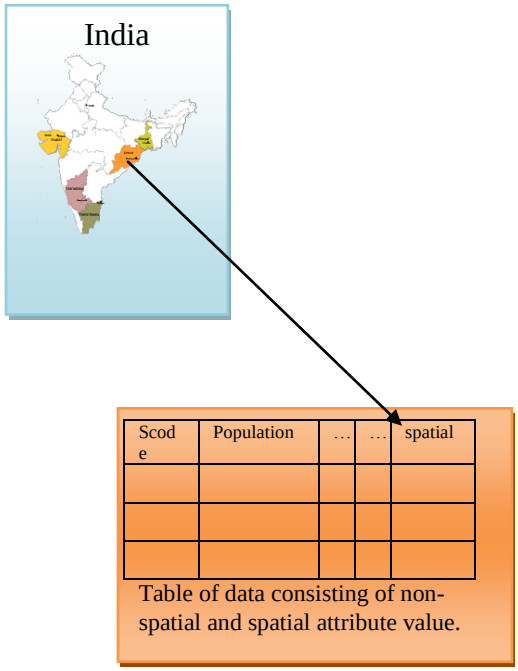


Figure 1: Non-spatial and spatial attribute values of Indian states.

The methods for discovering knowledge from spatial database focuses on non-spatial and/or spatial properties of spatial objects. Some important spatial data mining algorithms are:

2.3.1. Generalization-based methods for mining spatial characteristics and discriminate rules[4,6,28].

This is a widely used tuple-oriented technique in machine learning [13]. The method is often combined with generalization [14]. This approach cannot be used for large spatial database because the algorithms are exponential in the number of examples and it does not handle noise and inconsistent data very well. It requires the existence of background knowledge in the form of concept hierarchies. There can be two kinds of concept hierarchies *non-spatial* and *spatial* and are given by the experts or as per the requirement of the analysis. Following (Figure:2) could be a concept hierarchy of epidemiology study.

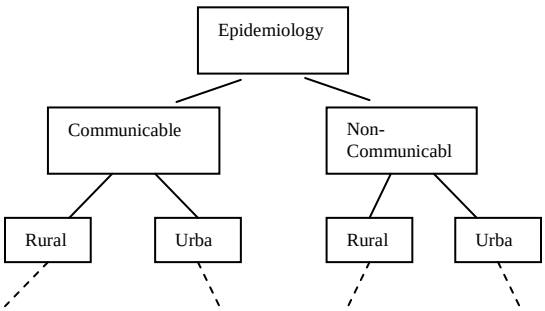


Figure 2: Concept Hierarchy of epidemiology

As we move upward in the concept hierarchy the information becomes more and more general. Similar concept hierarchy can be formed for spatial data for example hierarchy related to region, state, district, village etc. W. Lu and J. Han[6] described two generalization based algorithms one *spatial-data-dominant* and another *non-spatial-data-dominant* generalization. In the first approach the generalization of the spatial objects continues until the spatial *generalization threshold* is reached i.e. the no of region is not bigger than a threshold value. When the spatial-oriented induction process is complete, non-spatial data are retrieved and analyzed for each of the spatial object using the attribute oriented induction technique. The result of the query in spatial-data-dominant algorithm could be in the form of the follows map (Figure 3).

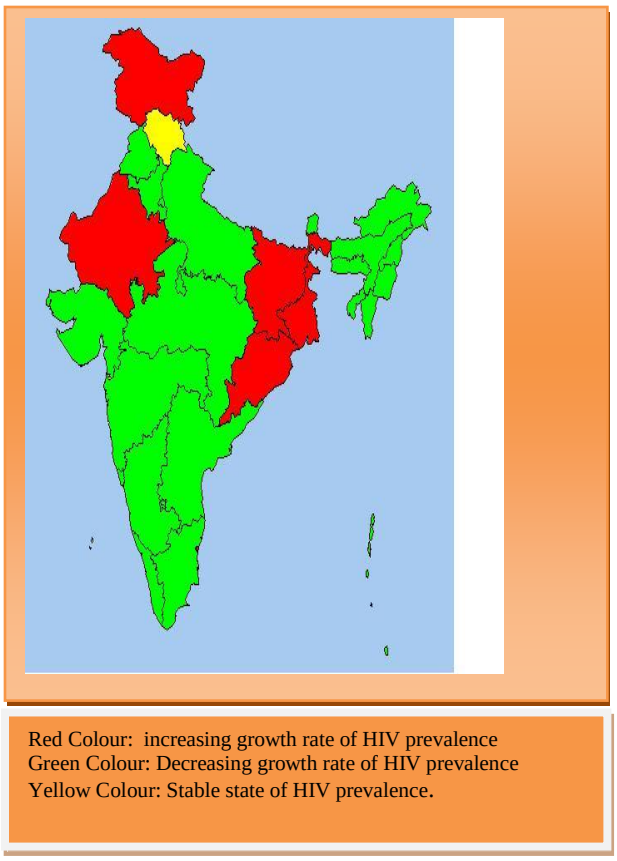


Figure 3 : The three categories of states having different growth rate of HIV. Output spatial data dominant.

In the second approach the algorithm performs attribute oriented induction on the non-spatial attributes, generalizing them to a higher concept level. The generalization threshold determines

whether to continue or stop the generalization process. In this process the pointer to the spatial objects are collected as a set and put with the generalized non-spatial data. Finally the neighboring area with the same generalized attributes are merged together based on the spatial function of adjacency (*adjacent_to*). For example the adjacent area having no of malaria epidemiology count ,both in male and female, more than 5% of population are merged together forming a *high-prevalence* cluster of malaria epidemiology. Similarly *low- prevalenc*e and *no-prevalence* clusters can be identified. The result of the query can be shown in the form of a map. An example of such a map (the dotted region inside the India map) is shown below Figure: 4 .

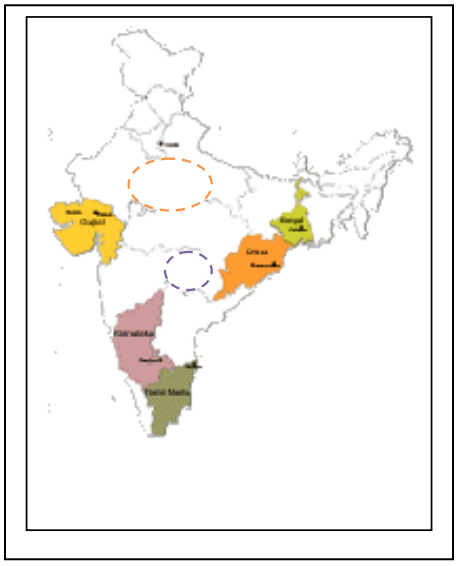


Figure 4 : Output non-spatial data dominant

In the above described generalization based algorithms, the concept hierarchy is generated automatically. However there are cases when such hierarchy is not present *a priory*. At the same time the mined characteristic rules is going to be much dependent on the given concept hierarchy(ies). Such concept hierarchy is given by the experts and therefore the above mentioned approaches falls under the category of *supervised classification or supervised knowledge discovery methods*.

On the other hand we have some *unsupervised techniques* of knowledge discovery methods. The method of clustering is one such approach. The conventional clustering algorithms like PAM (Partition Around Medoids) or CLARA (Clustering LARge Applications) [10], are not appropriate from computational complexity point of view. The difference between these two algorithms is that CLARA algorithm is based on sampling. CLARA can deal with large data set than PAM. Both PAM and CLARA were developed by Kaufman and Rousseeuw [10]. In PAM the cost of single iteration is $O(k(n-k)^2)$. Here n is the no of objects and k is no. of cluster. In CLARA the complexity of each iteration is $O(kS^2+k(n-k))$. Here S is the size of the sample.

Then CLARANS was developed for cluster analysis and it outperformed the previous two algorithms. This algorithm was proposed by Ng and Han [28] which tries to mix both PAM and CLARA by searching only the subset of data set and it does not confirm itself to any sample at any given time. Experimentally it has been shown that CLARANS is more efficient than PAM and CLARA. Its every iteration computational complexity is linearly proportional to number of objects [27]. Some of the drawback of CLARANS has been pointed out by Ester, Kriegel, and Xu[27]. It assumes that the objects to be

clustered are stored in main memory. For a large database it is not possible and hence a disk based method would be required. This method has been shorted out by integrating CLARANS with efficient spatial access methods, like R*-tree. But the construction of R*-tree is time consuming. Zhang, Ramakrishnan and Livny[29] presented another method BIRCH (Balanced Iterative Reducing and Clustering) for clustering of large set of points. The method is incremental one with possibility of adjustment of memory requirements to the size of memory that is available. It uses the concept called *Clustering Feature* and *CF tree*.

2.3.2. Two-step spatial computation technique for mining spatial association rules [34]

To minimize the number of costly spatial computation the two-step spatial computation technique for optimization during the search for association was introduced. Spatial association rule is a rule that associate one or more spatial object with other spatial objects. Agarwal, Imielinski and Swami [30] introduced the concept of *association rules* in the study of mining large transaction database. Later Koperski and Han[7] extended this concept to spatial database. In order to discover the useful rule the concept of *minimum support* and *minimum confidence* are used. A strong rule is a rule having large support and large confidence.

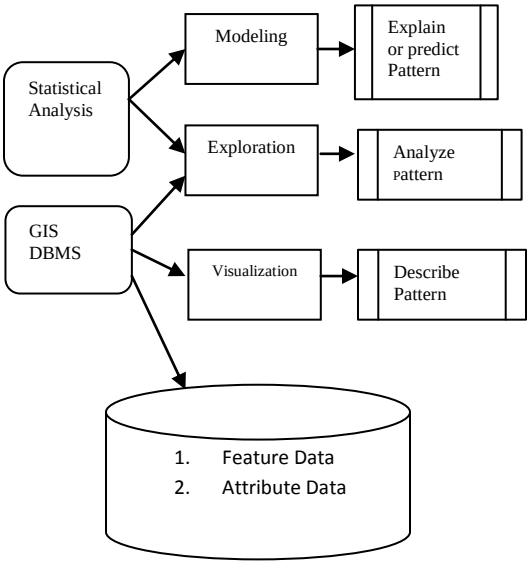
2.3.3. Aggregate proximity technique for finding characteristics of spatial clusters [31].

An aggregate proximity is the measure of closeness of the set of points in the cluster to a feature as opposed to the distance between a cluster boundary and the boundary of a feature. Related to a cluster it would be more interesting result to know why the clusters are there. The question that would more suitable answer about the cluster is that "what are the characteristics of the clusters in terms of the feature that are close to them". For example the statement like *85% of the houses in a cluster is close to the feature F (e.g. infected by infectious disease cholera)* would be more informative and interesting than statement like *one house is close to the feature F.*

3. Spatial Epidemiology- An Applicative Area

Elliott and Wartenberg [37] described "Spatial epidemiology is the description and analysis of geographic, or spatial, variations in disease with respect to demographic, environmental, behavioral, socioeconomic, genetic, and infectious risk factors". The spread of infectious disease is closely associated with the concepts of spatial and spatio-temporal proximity, as individuals who are linked in a spatial and temporal sense are at a high risk of getting infected [38]. Proximity to environmental risk factors is therefore important. Thus knowledge of spatial and temporal variations of disease and characterizing its spatial structure is essential for the epidemiologist to understand better the population's interactions with its environment [39].

Spatial epidemiology analysis comprises of wide range of methods. Now it is a big challenge to determine which one to use[38]. The figure below (Figure: 5) is a diagrammatic representation of a spatial analysis framework taken from Pfeiffer [38] adopted from Bailey and Bailey & Gatrell[40].



Source: Pfeiffer [38], Bailey and Gatrell [40]

Figure 5 : Conceptual framework of spatial epidemiological data analysis

In the above diagram Pfeiffer identified the following four active groups of the framework:

3.1.Data

Data is the basic need of epidemiological analysis which is conducted for description of spatial patterns, identification of disease cluster, and explanation or prediction of disease risk [38]. Geographic data system includes georeferenced feature data and attributes, be they point and area. These data are obtained by taking field survey, remotely sensed imagery or use of existing data generated either by government organizations or those closely linked to government such as cadastral, meteorological or national census statistics and health organizations.

3.2. GIS and DBMS

Management of the data is performed using GIS and database management system(DBMS), and is of relevance throughout the various phases of spatial data analysis. GIS provide a platform for managing these data, computing spatial relationship such as proximity to source of infection, connectivity and directional relationships between spatial units, and visualizing both the row data and results from spatial analysis within a cartographic context [38].

3.3. Visualization and exploration

It covers technique that focus solely on examining the spatial dimension of the data. Visualization tools are used resulting in maps that describe spatial patterns and which are useful for both stimulating more complex analysis and for communicating the results of such analysis. Exploration of spatial data involves the use of statistical methods to determine whether observed patterns are random in space. However there is some overlap between visualization and exploration, since meaningful visual presentation will require the use of quantitative analytical methods [41].

3.4. Modeling

Modeling introduces the concept of cause-effect relationships using both spatial and non-spatial data sources to explain or predict spatial patterns [38].

4. Future Direction

Data mining is a young field of study started during late 1980s. Spatial data mining is an even younger. The traditional data mining researchers extended their study to work on spatial data mining. Many spatial data mining methods assume the presence of extended relational model for spatial database. Some of the future directions of spatial data mining are enlisted below.

***Data Mining in Spatial Object-Oriented Databases*:** Many researchers have pointed out that OO database may be a better choice for handling spatial data rather than traditional relational or extended relational models[32,33].

***Mining Under Uncertainty*:** The use of evidential reasoning [34] can be explored in the mining process for the databases where uncertainty modeling has to be done. Bell, Anand and Shapcott [35] has explained that evidential theory can model uncertainty better than traditional probabilistic models, like Bayesian methods. Fuzzy sets approach was applied to spatial reasoning[20,36] and it can be extended to spatial data mining.

Mining Spatial Data Deviations and Evolution Rules: It is a more challenging and applicative work in spatial data mining. The work would be related to spatio-temporal databases to study data deviation and evolution rules. Foe example we can find spatial characteristic evolution rules which summarizes the general characteristics of the changing data. During the mining process we can discover the region having particular epidemiology growth rate more than the country's average growth rate. Similarly one can make a comparison of the areas where certain epidemiology increased last year with the area where it has decreased.

These rules may be used by the government and policy makers in formulating policies and plan to curb the problem.

Multidimensional Data Analysis and Rule Visualization: Discovering rule from multidimensional data (non-spatial and spatial) source is a challenge for the researchers. Multidimensional data analysis and visualization has been studied [42], but multidimensional rule visualization is still an immature area.

5. Conclusion

We have explained that spatial data mining is a promising field of research with wide application in GIS, medical and environmental data analysis etc. We surveyed the existing methods of spatial data mining and presented their strength and weaknesses. We have outlined one of the applicative area i.e. spatial data mining of epidemiology database which is of great importance for the society and policy makers and we hope to give some novel and useful output from our further exploration of this field.

6. Bibliography

[1] S.Shekhar and S.Chawla. Spatial Databases: A Tour. Pretice Hall (ISBN 0-7484-0064-6), 2003.

[2] R. Agarwal and R. Srikant. *Fast Algorithm for mining association rules*. In Proc. 1994 Int. Conf.

VLDB, pp. 487-499, Santiago, Chile, Sept. 1994.

[3] U. M. Fayyad, G Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Menlo Park, CA, 1996.

[4] J. Han, Y. Cai, and N. Cercone. *Data-Driven Discovery of Quantitative Rules in Relational Databases.* IEEE Trans. Knowledge and Data Eng., 5:29-40, 1993.

[5] G. Piatetsky-Shapiro and W. J. Frawley, editors. *Knowledge Discovery in Databases*. AAAI/MIT Press, Menlo Park, CA, 1991.

[6] W. Lu, J. Han, and B.C. Ooi. Discovery of General Knowledge in Large Spatial Databases. In Proc. For East Workshop on Geographic Information Systems pp. 275-289, Singapore, June 1993.

[7] K. Koperski and J. Han. Discovery of Spatial Association Rules in Geographic Information Databases. In Proc. 4th Int'l Symp. On large spatial Databases(SSD '95), pp. 47-66, Portland, Maine, August 1995.

[8] D. K. Y. Chiu, A. K. C. Wong, and B Cheung. A Statistical technique for Extracting Classificatory Knowledge from Databases. In Piatetsky-Shapiro and Frawley [43], pp 125-141.

[9] S. Fotheringham and P. Rogerson. *Spatial Analysis and GIS*, Taylor and Fransis, 1994.

[10] L. Kaufman and P J Rousseeuw. Finding groups in Data: an introduction to Cluster Analysis. John Wiley & Sons, 1990.

[11] G. Say, D. Wheeler, Statistical Techniques in Geographical Analysis. London, David Fulton, 1994.

[12] D. Fisher, Improving Interface through Conceptual Clustering. In Proc. 1987 AAAI Conf., pp. 461-465, Seattle, Washington, July 1987.

[13] R. S. Michalski, J. M. Carbonnel, and T. M. Mitchell, editors. Machine Learning: An Artificial Intelligence Approach. Morgan Kaufmann, Los Altos, CA, 1983.

[14] T. M. Mitchell. Generalization and Search. In Artificial Intelligence, 18:203-226, 1982.

[15] M. Stonebraker. *Reading in Database System*. Morgan Kaufmann, 1988.

[16] M. Stonebraker. *Reading in Database System*. 2ed.. Morgan Kaufmann, 1993.

[17] R. H. Guting. *An Introduction to Spatial Database System*. In VLDB Journal, 3(4):357-400, October 1994.

[18] R. Guttman. A dynamic index structure for spatial searching. In Proc. ACM SIGMOD Int. Conf. on Management of Data. Bostan, MA, 1984, pp. 47-57.

[19] H. Samet. *The Design and Analysis of Spatial Data Structure*. Addison-Wesley, 1990.

[20] S. Dutta. Qualitative Spatial Reasoning: *A Semi Quantitative Approach Using Fuzzy Logic*. In Proc. 1st Symp. SSD'89, pp. 345-364, Santa Barbara, CA, July 1989.

[21] M. J. Egenhofer. *Reasoning about Binary Topological Relation*. In Proc. 2nd Symp.

SSD'91, pp. 143-160, Zurich, Switzerland, August 1991.

[22] W. G. Aref and H. Samet . *Extending DBMS with Spatial operation*. In Proc 2nd Symp. SSD'91, pp. 299-318, Zurich, Switzerland, August 1991.

[23] W. G. Aref and H. Samet. *Optimization Strategies for Spatial Query Processing*. In Proc. 17th Int. Conf. VLDB, pp. 81-90, Barcelona, Spain, Sept. 1991.

[24] J. Han, and Y. Fu. *Exploration of the power of Attribute-Oriented Induction in Data Mining*. In[16]

[25] M. Holsheimer and M. Kersten. Architectural Support for Data Mining. In CWI Technical Report CS-R9429, Amsterdam, The Netherlands, 1994.

[26] C. J. Matheus, P. K. Chan, and G. Piatetsky-Shapiro. *Systems for Knowledge Discovery in Databases.* In IEEE Trans. Knowledge and Data Engineering, 5:903-913,1993.

[27] M. Ester, H.–P. Kriegel, and X. Xu. *Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification*. In Proc. 4th Int. Symp. On Large Spatial Databases (SSD'95),pp.67-82, Portland, Maine, August 1995.

[28] R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. In Proc. 1994 Int. Conf. Very Large Databases, pp. 144-155, Santiago, Chile, September 1994.

[29] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an Efficient Data Clustering Method for Very Large Databases. In Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data, Montreal, Canada, June 1996.

[30] R. Agarwal, T. Imielinski, and A. Swami. *Mining Association Rules Between Sets of Items in Large Databases.* In Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data, pp. 207-216, Washington, D.C., May 1993.

[31] E. Knorr and R. T. Ng. *Applying Computational Geometry Concepts to Discovering Spatial Aggregate, Proximity Relationships.* In Technical Report,University of British Columbia,1995.

[32] L. Mohan and R. L. Kashyap. *An Object-Oriented Knowledge Representation for Spatial Information.* In IEEE Transaction on Software Engineering, 5:675-681, May 1988.

[33] J. Han, S. Nishio, and H. Kawano. Knowledge Discovery in Object-Oriented and Active Databases. In F. Fuchi and T. Yokoi(eds), Knowledge Building and Knowledge Sharing, Ohmsha/IOS Press, pp. 221-230, 1994.

[34] J. Guan and D. Bell. Evidence Theory and its Applications, vol. 1. North-Holland, 1991.

[35] D. A. Bell, S. S. Anand, and C. M. Shapcott. *Database Mining in Spatial Databases.* International Workshop on Spati-Temporal Databases,1994.

[36] S. Dutta. Topological Constraints: *A Representational Framework for approximate Spatial and Temporal Reasoning*. In Proc. 2nd Symp. SSD'91,

pp.161-182, Zurich, Switzerland, August 1991.

[37] P. Elliott and D. Wartenberg. Spatial epidemiology: current approaches and future challenges. Environmental health perspectives, 112(9):998, 2004.

[38] Dirk Pfeiffer. Spatial analysis in epidemiology. Oxford University Press, GB, 2008.

[39] Frank B. Osei. *Spatial statistics of epidemic data* : the case of cholera epidemiology in Ghana. PhD thesis, 2010.

[40] T.C. Bailey and A.C. Gatrell. Interactive spatial data analysis. Longman Scientific & Technical Essex, 1995.

[41] A. Maroko, J.A. Maantay, and K. Grady. Using geovisualization and geospatial analysis to explore respiratory disease and environmental health justice in New York city. Geospatial Analysis of Environmental Health, pages 39–66, 2011.

[42] D. Keim, H. P. Kriegel, and T. Seidl. Supporting Data Mining of Large Database by Visual Feedback Queries In Proc. 10th of Int. Conf. on Data Engineering, Houston, TX, pp. 302-313, Feb. 1994.