Intelligent Knowledge Discovery in Spatial Data Sets

[1]Arvind Sharma, [2]R.K.Gupta
[1]RJIT, Tekanpur, Madhya Pradesh, India
[2] MITS, Gwalior, Madhya Pradesh, India

Abstract **-** Spatial databases are increasingly used to set trends, models, by providing input parameters for spatially explicit process simulators. Global geographic, Biochemical, Meterological models, for example are based on parameter maps of multiple spatially distributed variables like age density, location, land cover, cost, temperature etc. are typically provided in the form of trends, graphs, information layers in spatial data mining process. Arc viewer, Arc Catalog, Arc GIS, Map Calc etc simulators and software are helpful for designing spatial data mining models and also set trends for prediction spatial characteristics on the basis of indirect collection of parameter's values. Such a synthesis involves generation and prediction of parameter values at random and un sampled locations and calls the appropriate algorithm and queries to obtain parameter estimates on a usually regular grid. Present paper gives the overview of geographic and applied research in spatial data mining and its some common tasks like clustering, classification, regression, interpolation etc. The articles and results involved in this paper contribute to geo analysis, map analysis and development of new algorithms for optimization of spatial data mining tasks.

Keywords **-** Spatial Data Mining (SPDM), ArcCatalog, geo visualisation, geo databases.

I. INTRODUCTION

It has observed and revealed by different scientist and researcher that some fields of SPDM and geographic research are observational rather than experimental. They also observed that spatial databases, geographic data and related problems are too complex. With the rigorous work of scientific projects, government agencies, and private sectors, huge geographic data have been generated and collected. We now can generate ,obtain much more diverse, dynamic, realtime, synthetic data with modern data collection and generation schemes like GISArcview, ArcCatalog, MapCalc,high resolution remote sensing and many more. The availability of mass and high resolution data provides better opportunities to understand most hidden and novel facts of spatial data bases such as understanding of social, economic, dynamics,Human environment interaction,which are related with real world problems[2].These application areas may be weather forecasting,land distribution with costing,density,livingstandards,spread of diseases in particular areas ,global environmental changes and many more.

There are so many forms and types of spatial databases are available but in this paper we have applied special nature of spatial data bases which are generated and collected from top most agencies and sources of spatial data bases.Our model supports following types of nature of data as .shp,.tab,.dbf,.csv,.txt,.xls,.rgs,.mif,.tgt,.grdetc. and contain all basic requirements of SPDM supporting parameters and nature. These files may be generated or created as per requirement by application of certain software tools like ArcView,ArcCatalog,GISView,MapCalc etc.

Our experimental results are based on MapCalc software tool that is specially designed for SPDM .When we go in deep at comparative state then found other GIS software supporting surface modelling,geo statistics and spatial analysis has been too expensive and complicated for all advanced scientific and research work. The MapCalc package is a set of inexpensive functions ,graphic based materials for learning map analysis concepts ,considerations and procedures.it provides basic set of exercises and data sets withall possible functions which are necessary for understanding of process of Spatial Data Mining in graphical manner[1].This paper describes and demonstrates some important visual analysis of spatial data bases and mining which are core part of this software.This paper contains several real time demonstration that are directly able to unfold various new facts of spatial data mining process.Vector based systems are ideal for learning the fundamentals of mapping and spatial database management[12].The way and exposure of results are based on grid based map analysis.A major feature of this software is inexpensive and user friendly desktop mapping functions.

II. MAP ANALYSIS THEORY AND ANALYSIS FRAME STRUCTURE

There are threemajor map features[12] that comprise all kind of maps like –points,line,andpolygons.There is a choice that if we consider or apply grid based analysis the another important map feature plays a very important role and this is **a surface**.[4]The grid based construct enables display and processing of geographic space as a continuum.The base spatial unit is a grid cell and is used to identify –

- Points – Single cells
- Lines - connected cells of sets
- Polygons - All cells identifying the edge and interior of the parcel,and
- Surface – All cells within a project area with a value assigned to each that indicates the presence by feature type(discrete object)or the relevant variable response(continuous gradient).

The figure 1 shows geographic space as a continuum[12].

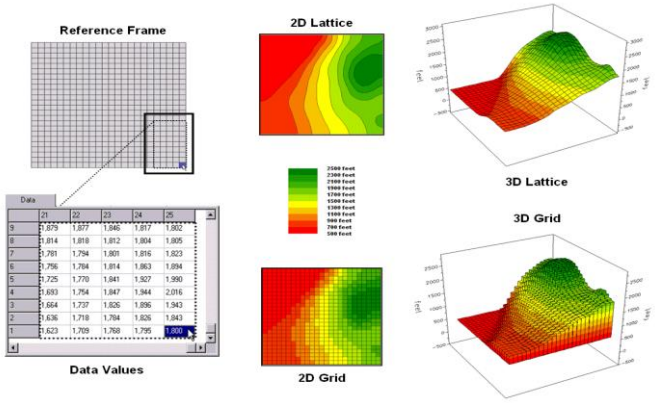


Fig.1: Grid-based data can be displayed in 2D/3D lattice or grid forms.

Upper and lower portions are shown in 2D and 3D gridmap, whenever we want a lattice structure then simply connect the grid lines at the centre of each grid space. Different values are shown by different lengths of lines. The results are shown in the form of 'wire frame' that forms the peaks and valleys of the spatial distribution of the data. As we have told that our software contain so many functions and parameter based queries to make new and innovative patterns with the application of colour codes. Shadowing property shows grid structure of same data.3D form of figure forms a stepped surface with a single value across the entire cell. In further steps it is applicable for deep map analysis.

III. PRACTICAL APPROACH OF GEOGRAPHIC DATA ANALYSIS IN TERMS OF MAP ANALYSIS

Simple procedure to map analysis starts from division of map into a uniform set of grid cells (pareels). [6]The result of analysis provides a sort of geographic consistency for setting of spatial relationships within and among grid layers.

It has been observed by different practical results that vector systems are best suited for computer mapping and geo – query but have limited map analysis capacity.[13]Grid systems are lesser competent in computer mapping and query but contain a robust set of analytical operations. For complete analysis of map we perform slope map analysis and surface flow of map from the generated figures and operations. Given in figure 2.This figure shows 2D and 3D views of slope map derived using the analysis frame.

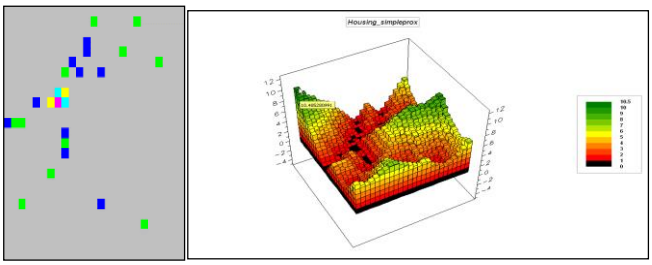


Fig.2: Map and its connectivity in 3D form

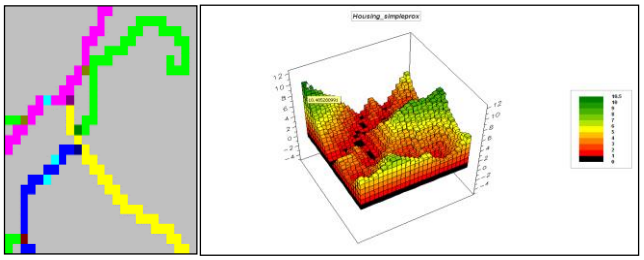


Fig.3: View after performing operation

The larger 3D display shows the slope map draped over the elevation surface. It is noteworthy that the steep areas (green) and flat areas (red) align with the appropriate surface slopes or inclinations providing visual conformation of the calculated slope values.[12]moving of 3*3 window over the entire elevation surface gives best fit of elevation values in the roving window. On the other hand surface flow map, areas with higher flow values (green)align with the small ravines visible on the terrain surface.[2]The process simulated a drop of rain feeling at each grid cell, tracing its steepest downhill path while accumulating the number of paths that cross each cell, higher number indicate locations of water confluence.

IV. WORKING MODEL AND ITS KEY CONCEPT

The result and expected values of every map analysis depends on type of proposed model[12] and its computed values.[5] Sequencing operations for develop analysis procedures, such as erosion potential model are shown here-

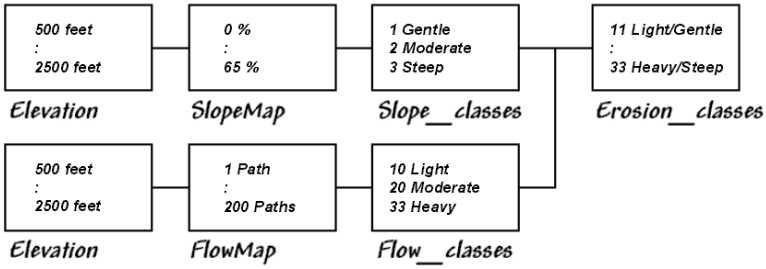


Fig.4: Flow chart model for estimating soil erosion

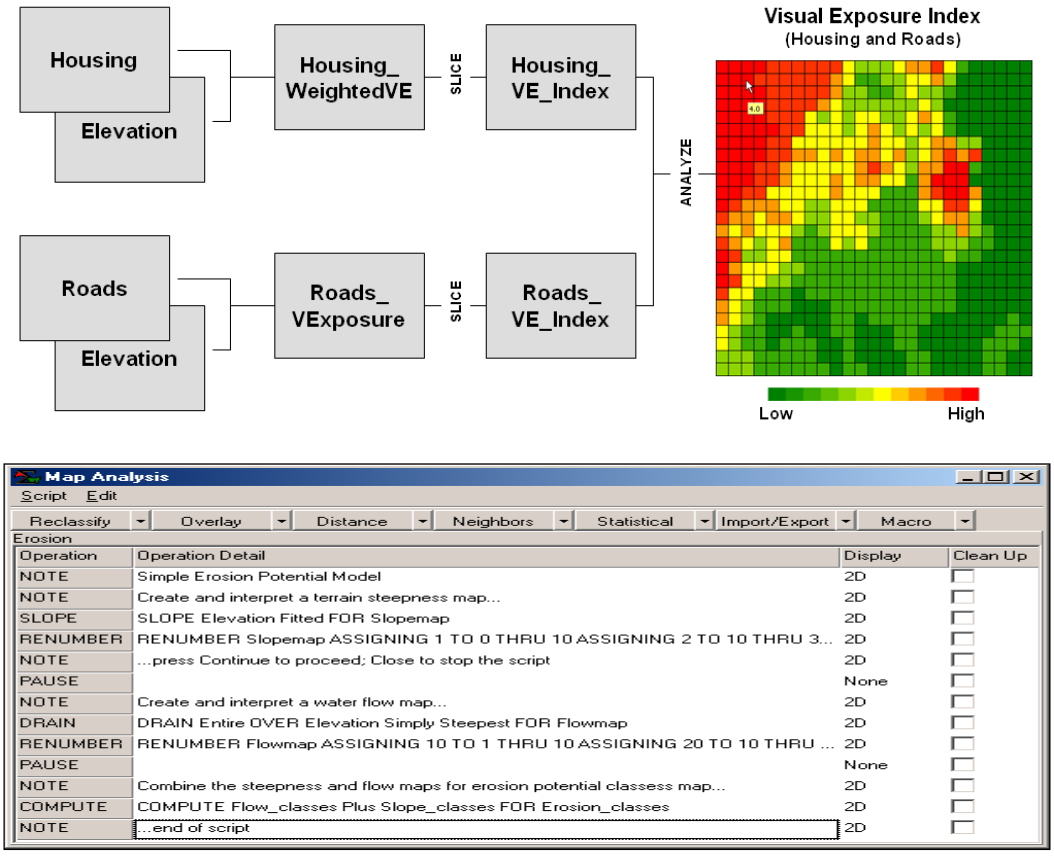


Fig.5: Snapshot of available functions and command sequence.

After selection of a specific spatial data type and performing various map analysis operations as per given list of operations in the software, computed values as slope and flow may be depicted by figure- 6 and combination of two maps construct a simple erosion potential model[6].

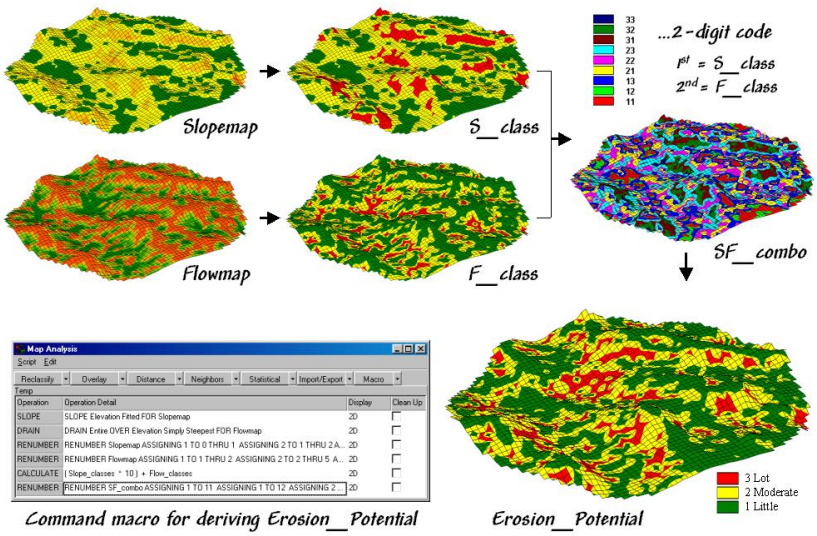


Fig.6: An example applying the *Erosion Potential Model*

Command like SLOPE, DRAIN and RENUMBER, CLASSIFY, and COMPUTE operations are necessary for simple erosion model. Application of further commands like SPREAD operation is used to calculate effective distance from the streams and lakes on the water map. Map analysis operations are listed here. All operations are 26 and grouped into five classes – Reclassify, Overlay, Distance, Neighbors, and statistics as shown in figure-4.Contextual help is available for all commands [2].

MapCalc Command	Grid / Spatial Analyst
Slope*creates a map indicating the slope (1st derivative) along a continuous surface.*	*Surface function* ***SLOPE***
Drain*creates a map indicating the number of steepest paths (optimal path density) from a set of locations along a surface.*	*Hydrologic function* ***FLOWACCUMULATION***
Renumber*assigns new values to the categories on a map.*	*Reclassification function* ***RECLASS***
Calculate *creates a map as the mathematical or statistical function evaluating an equation using a stack of map layers.*	*Arithmetic operators* ****, +, -, DIV, MOD***
Composite *creates a map summarizing values from one map that coincide with the categories of another.*	*Zonal functions* ***ZONALFILL, ZONALMAX, ZONAL MEAN, ZONALMIN, ZONALRANGE, ZONALSTD, ZONALSUM, ZONALVARIETY***
Spread*creates a map indicating the shortest effective distance from specified cells to all other locations.*	*Distance functions* ***CORRIDOR (compute sum), COSTALLOCATION (slice), COSTDISTANCE, EUCALLOCATION, EUDIRECTION (orient), EUCDISTANCE*** *Shape Analysis functions* ***EXPAND, SHRINK*** *Hydrologic function* ***WATERSHED, BASIN***
Radiate*creates a map indicating areas that are visible from specified locations.*	*Visibility tools* ***VISENCODE, VISIBILITY***
Cover *creates a new map where non-zero values of the top map replace the values on the previous (bottom) map, or stack of maps.*	*Selection functions* ***SELECTBOX, SELECTCIRCLE, SELECTMASK, SELECTPOINT, SELECTPOLYGON***

Table 1: List of map analysis operations in the software.

V. DATA COLLECTION AND DATA GENERATION

Spatial data may be collected from different web sites or generated from the software. A sample is given below which is collected from a spatial data base port [8, 9, 13, 14].

id	Bus Stop Name	Street	Location	Cross Street	Suburb	Latitude	Longitude
1	Joseph Banks Ave	Joseph Banks Av	App	LeichhardtCct	FOREST LAKE	-27.6137	152.969
2	Forest Lake B	Forest Lake Bvd	FS	Joseph Banks Av	FOREST LAKE	-27.6155	152.9651
3	Bus Stop A	Forest Lake Bvd	FS	Settlers Cct	FOREST LAKE	-27.6191	152.9659
4		Forest Lake Bvd	App	High St	FOREST LAKE	-27.6238	152.9702
5	Cremin Street 45/60	Kessels Rd	App	Cremin St	UPPER MOUNT GRAVATT	-27.5603	153.0797
6		Nagle St	App	Bolger St	UPPER MOUNT GRAVATT	-27.5584	153.0771
7		Nagle St	App	Pickworth St	UPPER MOUNT GRAVATT	-27.5582	153.0757
8		Meckiff St	Opp	Sobers St	UPPER MOUNT GRAVATT	-27.5578	153.0737
9	Tora Street - 77	Tora St	App	Ilya St	MACGREGOR	-27.5565	153.0711
10	Omeo Street - 76	Omeo St	App	Ilya St	MACGREGOR	-27.5579	153.0687
						-27.561	153.0655
						-27.6196	152.9722

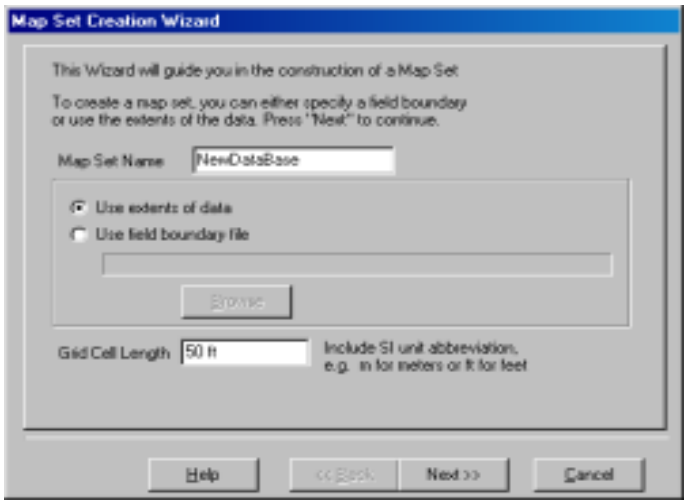
Another sample is given as[11][13][14] –

id	Nursery Name	Address	Suburb	Phone Number	Latitude	Longitude
1	Carseldine Nursery	769 Beams Road	Carseldine	07 3263 2504	-27.34498	153.00724
2	Edible Landscapes Nursery	16 Victoria Street	Windsor	07 3857 8774	-27.442659	153.029221
3	Nova Gardens Nursery	78a Settlement Road	The Gap	07 3300 4161	-27.43883	152.9434
4	King Avenue Cultivators	46 King Avenue	Willawong	07 3879 0222	-27.59409	153.00235
5	Cottage Garden Nursery	999 Stanley Street East	East Brisbane	07 3891 7999	-27.48825	153.04874
6	Hawkins Garden Centre	1666 Old Cleveland Road	Chandler	07 3390 1633	-27.5086	153.14579
7	Ross Evans Garden Centre	2274 Moggill Road	Kenmore	07 3378 5553	-27.5172	152.92358
8	Greenworld Plants Plus	22 Gowan Road	Sunnybank Hills	07 3344 5121	-27.591028	153.065277

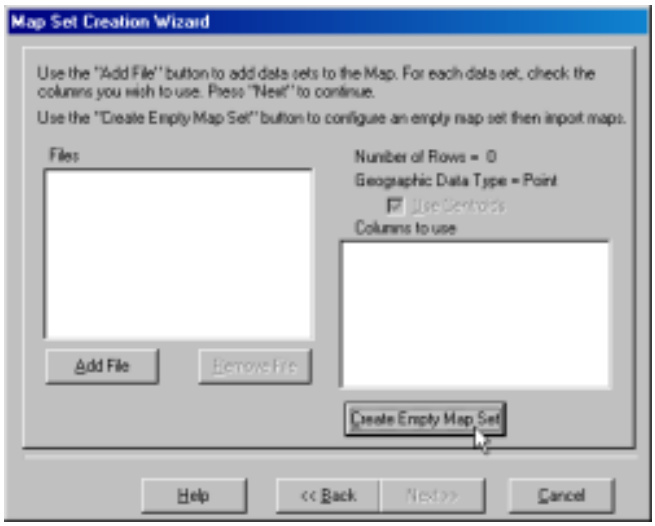
And finally another process of data generation which is actually based on geographic generation of data sets has shown here with the help of given software. Our experiments are based on

Creating Your Own Database (Empty Map Set)

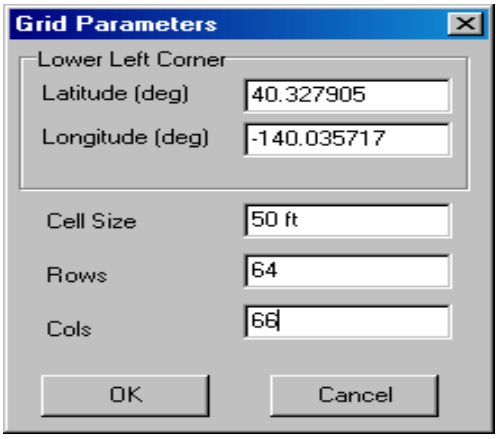
Click on the Create a new file button or from the main MapCalc menu, select File New [12].



Specify a name for your new data base (e.g., **NewDataBase**) then press next -



Click on **Create Empty Map Set** button to pop-up the Grid Parameters dialog box.



Specify the **Latitude** and **Longitude** of the lower-left corner of the analysis grid (e.g., Lat= **40.327905**, Lon= **-140.035717**). Enter the **Cell Size** (e.g., **50** feet) and the number of **Rows** and **Columns** comprising the analysis grid area (e.g., **64** rows and **66** columns). Click **OK** to create the empty database.

Here we have shown both possibilities of data collection and generation.

VI. FUTURE SCOPE

Application of spatial data mining approach for spatial object oriented data sets, multidimensional data sets, and identification of similar objects or patterns with some certain additional techniques like fuzzy technique or machine learning intelligent methods are still considerable. Spatial data mining is very large area with large data bases for further enhancement of research activities so it is always needed to apply the concept of parallel computing to satisfy the results according to demand and complexity of data.

VII. CONCLUSION

On the basis of different data sets which may be collected from different data sources of different locations and all these data sets come in the category of spatial data sets. As described in the beginning that all files which belong to .shp,.csv,.rgs,.txt,.xls and many more are the real world examples of spatial data sets. Analysis of these data sets on our software has been completed successfully. Basic requirement of data sets are that they should be compatible with attribute of spatial data types or sets. These data sets [13]may contain health data, Finanace data, Cultural data, Locations data, medicine data, satellite images, weather or climate data, maps of space, buildings, areas, growths etc.

After performing analysis on different data sets of different locations, certain trends [13] may be set fot further prediction. Special trends describe a regular change of nonspatial attributes when moving away from certain objects. Global and local trends can be distinguished to detect and explain such spatial trends e.g with respect to the weather forecasting system, is an important issue in geo graphic information system. Spatial characterization not

only consider the attributes of the target regions but also neighboring regions and their properties. Input file should have non spatial attributes as well as spatial attributes. Non spatial attributes are like name, population, rate of growth etc. and spatial attributes are coordinates, latitude, longitude, elevation as well as shape etc. In a nut shell this is to say that this paper presents the techniques of spatial data mining in the following categories – clustering and outlier detection, association and co-location, classification, trend detection and map analysis. Finally with the help of improved algorithms, results of map analysis and procedure of treatment of spatial data in better way may be increased.

VIII. REFERENCES

[1]. Anselin L. 2008, Exploratory Spatial Data Analysis and Geographic information systems.InpainLo,M.Ed.,New tools for SpataialAnalysisd ,45-54.

[2]. Analyzing Geo business data "Joseph K. berry,2003"

[3]. Bending our understanding of distance ;calculate effective data distance and connectivity by Joseph k. Berry.

[4]. [Anselin 1988],Anselin L. 1988,Spatial econometrics:Methods and Models .Dordrecht,Netherlands:Kluwer.

[5]. [Han, Kember,& Tung 2001]Han.J .Kamber, M;and Tung, A.2001.Spatial clustering methodsin Data mining.

[6]. [Morimoto 2001] Morimoto ,Y. 2001. Minimum frequency neighboring class sets in spatial data bases.In Proc. ACM SIGKDD International conference on knowledge discovery and data mining.

[7]. [Shekhar at el 2004] Sekhar, S. Schrator , P.R; Vatsvai, R.R; Wu,W;and Chawla, S;2002. Spatial contextual classification and prediction models for mining Geo Spatial data.

[8]. [software.esri.com]

[9]. [ftp2.cites.rncan.gc.ca]

[10]. www.kdnuggests.com

[11]. www.geodata.gov.gr

[12]. www.innovativegis.com

[13]. www.jcaksrce.org

[14]. Data.gov.au

[15]. S. Shekhar, R. Vatsavai, Techniques for Mining Geospatial Databases, as Chapter 22 in Handbook of Data Mining, Nong Ye (Eds.), LEA Publishers, NJ, 2003.

[16]. S. Shekhar, C.T. Lu, P. Zhang Detecting Graph-based Spatial Outliers: Algorithms and Applications, *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001, San Francisco, CA, 2001*.