An Integrated Approach of GIS and Spatial Data Mining in Big Data

Hemlata Goyal
Research Scholar
Banasthali University

Chilka Sharma
Department of Remote Sensing
Banasthali University

Nisheeth Joshi
Department of Computer Science
Banasthali University

ABSTRACT
An explosive growth of spatial data has been demanding to Spatial Data Mining (SDM) technology, emerging as a innovative area for spatial data analysis. Geographical Information System (GIS) contains heterogeneous data from multidisciplinary sources in different formats. Geodatabase is the repository of GIS data, representing spatial attributes, with respect to location. Rapidly increasing satellite imagery and geodatabases generates huge data volume related to real world and natural resources such as soil, water, temperature, vegetation, forest cover etc. Inferring information from geodatabases has gained value using computational algorithms. The intent of this paper is to introduce with GIS, and spatial data mining, GIS and SDM tools, algorithmic approaches, issues and challenges, and role of spatial association rule mining in big data of GIS.

Keywords
GIS, SDM, Geodatabases, Spatial and Nonspatial data, Bigdata, MRPrePost

1. INTRODUCTION
GIS has advanced as a new emerging field as enhancement of communication technologies. The rapid growth of data, information, and communication has generated voluminous data of earth surface. Just of increasing powerful remote sensors, more computing power and enhancement in GIS technologies themselves, GIS is developing platform of selection for integrating and analyzing massive amount of earth data. Today's GIS is become indispensible and used in multidisciplinary areas to access information with respect to position. Enormous amount of GIS data is collected in numerical, text, graphics and analogues forms from satellite imagery sensors and other devices which represent the spatial and temporal situation. Spatial data mining has been emerging as innovative research area for data analysis with respect to spatial relations. SDM techniques has strong relationship with GIS and widely used in GIS for inferring association among spatial attributes, clustering, and classifying information with respect to spatial attributes. This paper gives the idea to understand GIS data models, data sets, data sources to provide better understanding of GIS for analyzing the data using spatial data mining techniques. This paper is organized as follows, section 2 with a detailed knowledge of GIS, data sources, data models. Section 3 provides with description of SDM tasks applied in various domains of GIS data. It also presents the overview of SDM tools for GIS. Section 4 describes the issues and challenges with respect to GIS data set and architecture and methodology is proposed for the same finally drawn by conclusion in section 5.

2. GIS
Geographical Information System-GIS acts as information systems for importing, storing, analyzing, managing, exporting and presenting spatial referenced data(linked to location) [1]. Because of the technological advancement in automated data acquisition in GIS domain has generated voluminous geographic data to represent spatial nomenclature of earth surface. GIS data received from heterogeneous discipline such as high resolution remote sensors, global positioning systems (GPS), location aware services and surveys etc. GIS database stored geo-reference spatially related dataset received from aforesaid heterogeneous components connected to each other and provide the spatial information about location, links with other, and description of nonspatial (attribute) features. Spatially located dataset gives the information about to what, when and where. GIS has powered significance in analysis related to knowledge management and data mining. It is a collection of components with respect to position namely Data, Software, Hardware, Procedures and Methods [2] for analysis and decision making as shown in Fig.1.

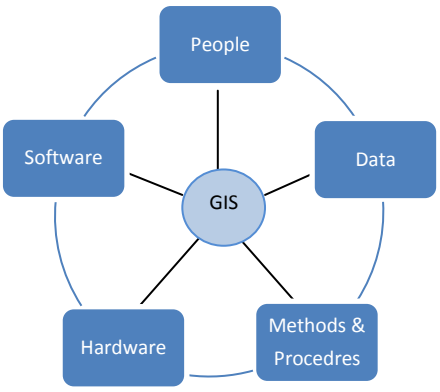

Fig.1. Component of GIS

Spatial Data Mining techniques collectively used with GIS and satellite imagery in various studies to mine interesting facts associated in diverse domains' applications such as traffic risk analysis, fire accident analysis, analysis of forest extent changes, grading of agriculture land, analysis of railways, farming and forestry, warehouse, transport, tourism, military, geology, soil quality monitoring, water resource monitoring, and deforestation, land allocation, meteorology [3-17]. This section throws the light on GIS data source, data formats, trends and applications in GIS.

2.1 GIS Data Sources
GIS dataset are of two types-spatial dataset and attribute dataset. Spatial data is about to where while attribute data is about to what. An entity contains the information about both spatial and attributes data. For example a spatial point entity is represented in GIS by georeference location using latitude and longitude and related to what type of information contain about that point, is described by nonspatial attribute data. The

geodatabase is used to store a collection of geographical datasets having three elements-space, theme, and time. The object is characterized in the form of line, point and polygon, pertaining to object in the database. GIS objects are uniquely identified by latitude and longitude. Spatial data can store, retrieve and manipulate with GIS. The various available dataset of GIS are listed in Table1 to summarize details.

Table 1 GIS Data Sources

Source of Data	Description
USGS Earth Explorer	Satellite image sensory data
Natural Earth	Cultural, physical and raster (basemap) data.
OpenStreetMap	Vector data of high spatial resolution vector data of cultural.
NASA's (SEDAC)	Remote sensing climate, sustainability, urban and water, land use agriculture etc.- Socioeconomic data
ISCGM Global Map	Landuse, landcover, elevation, vegetation etc.
Geocoder.us	Provides latitude & longitude of any US address , Geocoding for incomplete address ,Bulk Geocoding and Calculates distances
CityGrid	Incorporates local content into web and mobile applications.
Yahoo! BOSS	BOSS (Build your Own Search Service). Provides a facility of Place finder & Place Spotter to make location aware.
GeoNames	It contains geographical names, populated places and alternate names. All categorized into one out of nine feature classes.
ArcGis,MapGis	Helps to organize and analyze geographic data
MaxMind	GeoIP - IP Intelligence databases and web services minFraud-transaction fraud detection database
US Census	Offers several file types for mapping geographic data based on data found in our MAF/TIGER database
OpenEarthquake Data	Gives the databases related to earthquake

2.2 Data Modeling in GIS

GIS Systems deposits data from different varied data sources from wide range of communicating devices in different representation and file format. Representation of data can be done in two main categories as raster and vector data types, shown in Fig. 2.

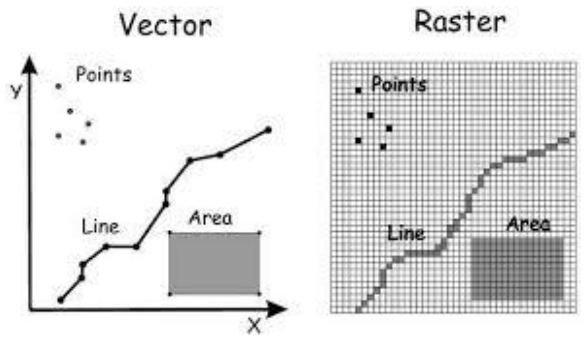

Fig.2. GIS Data Model (ref. [18])

Raster data type is a two dimensional grid of rows and columns and stores the information as the value of pixel colors in a cell and the attributes values are contiguous in nature. Raster data type is used to represents information about to real world objects precisely such as aerial photograph, scanned maps, remote sensing data. It is best suitable representation technique for two dimensional spatial entities such as line, area and network. Vector data types are used to represent discrete features in nature. It is better representation method for surface representation in GIS and has a layered architecture representing point, line, and polygon. Vector data types are used to represents information from sources such as roads, rivers, cities, lakes, park boundaries with a layered hierarchy. TIN, contour, interpolation, elevation etc features can be easily extracted using vector data model. Raster to vector conversion shown in Fig.3.

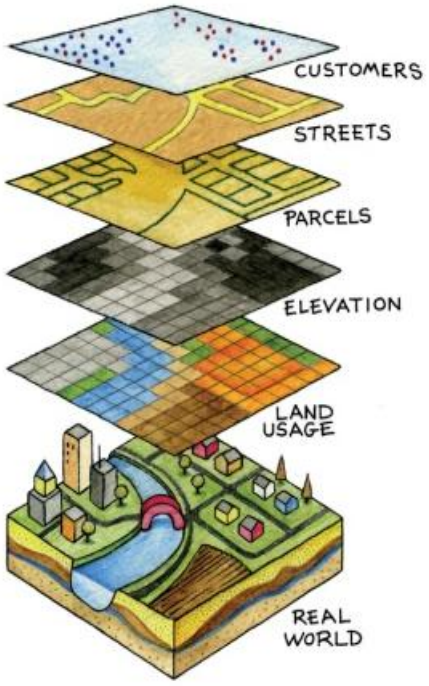

Fig.3. Layers-Raster to Vector (ref. [19])

2.3 GIS Challenges

Analysis of GIS databases representing spatial information is complex due to the varied formats, representation and data

sources. GIS has to face various challenges involved mining geospatial data from geodatabases as listed below:

- require help of field specialists to relate and comprehend Spatial and nonspatial Data.
- selection and representation of data for mining from geodatabases due to wide range of file formats.
- comprehension of data represented in image form(raster information).
- selection and transformation of spatial attributes from Non spatial attributes.
- false precision (obscuring sources of error)
- over-investment in data irrelevant to research goals or decision-making

3. SPATIAL DATA MINING

A large amount of spatial dataset is increasing and accumulated from remote satellite imagery, sensors, aerial photography, etc. and this voluminous data have exceeded beyond to human capability to fully modeling, interpreting, analyzing and using. Therefore some advanced and efficient technique is required to discover knowledge from huge spatial database [20]. Spatial Data Mining is the technique to find out the knowledge from huge geospatial dataset for extracting unknown, necessary spatial relationship, trends or patterns, not stored explicitly in spatial database [21]. SDM techniques is nothing new, just an expansion of data mining techniques applied on Spatial data, GIS data and satellite dataset associated in various domain. Spatial means the data associated with the geographic location of the earth. Spatial data mining has been often used in many applications, like remote sensing, Marine Ecology, space exploration, environmental science, resource management, agriculture, geology, climatic change, NASA Earth Observing System (EOS), traffic analysis, Census Bureau, National Inst. of Health, National Inst. of Justice, etc. there is a need of mining the frequent trajectory patterns in a spatial–temporal database [22], extracting the spatial association rules from a remotely sensed database, generating polygon data from heterogeneous spatial information, and analyzing the change of land use [23]. Spatial data mining is the extraction of spatial rules [21] and has been used in many real time applications. [24] proposed a model to that selects the locations of land-use by using the decision rules generated by nearest neighbours. It enables to recognize spatial dataset, find out relationship between spatial and nonspatial dataset, building of spatial knowledgebase, query optimizations, reorganizations of the spatial data, capture the basic features in easy and summarize way, etc. Spatial data using traditional data mining methods such as association, classification, clustering, trend detection, outliers generates interesting facts in associated domain. Spatial data mining techniques deals with spatial and nonspatial objects, attributes of neighboring objects and their spatial relations to find class, making clustering rules to detect outliers and deviations of trends, find association to extract multilevel topological relations. We will discuss these spatial data mining techniques in rest of the paper. Before this let us know about to spatial data and databases.

3.1 Spatial Data and Query

Spatial objects are uniquely identified by latitude and longitude with respect to geo-reference. Generally a GIS are used to store information related to geographic locations on the surface of the earth, retrieve and manipulate it. Spatial entities are used to store in the form of point, line, area, network and surface while modeling. Spatial data represents geographic coordinates in number format. It is multidimensional and auto correlated. It includes location, shape, size and orientation. Nonspatial data/attribute data is not dependent on geometric considerations. Nonspatial data includes height, mass and age, etc.

As the complexity of spatial operations, spatial query processing, its optimization needed much work to perform. Therefore, data accessing for spatial data is much difficult than nonspatial data. In traditional selection query for nonspatial data used standard comparison operations like >, <, <=, >=, not equal to. For spatial data that could be used as spatial comparator include east, west, north, south, near, contained in, and overlap or intersect. Spatial selection query, for e.g. find all schools near to downtown. In traditional join, two records must have attributes in common that satisfy a predefined relationship. In some ways, a spatial join is like a regular relational join in which two records are joined together if they have features in common. For example, the nearest relationship may be used for points, while the intersecting relationship is used for polygons. Some of the basic spatial query include: A region query is used to find the intersected in a given region; A nearest neighbour query find objects which are close to an given object; A distance scan finds objects within a certain distance of an given object. In general the spatial attributes are classified in three major relations as distance relation, direction relation and topological relation. Topological relation is always non spatial data, so it requires spatial mapping to convert non spatial to spatial data [25]. The advance in spatial data has enabled efficient querying of large spatial databases using spatial operations as shown in Table 2.

Table 2 Data model and Operations [26]

Relationship-Nonspatial	Relationship-Spatial
Arithmetic	Set oriented-Union, intersection, membership
Ordering	Topological-Meet, within, overlap
Subclass-of	Metric-Distance, area, destroy
Isinstance-of	Direction-North, NE, above, left, behind
Membership-of	Network-Shape-based and visibility
Part-of	Dynamic-Update, create and destroy

3.2 Spatial Data Mining Operations

Operations needed to support spatial data mining involve those required for spatial databases. As defined in [27], there are several topological relationships that can exist between two spatial objects A and B in 2-d space. Consisting of a set of points in the space, each object can be viewed as: (x_a, y_a) belongs to A and (x_b, y_b) belongs to B. These two objects are placed in geographic space are based on disjoint, overlaps or intersect, equals, contained in, contains.

3.2.1 Disjoint

A is disjoint from B if there are no points in A that are contained in B.

3.2.2 Overlaps or Intersects

A overlap with B if there is at least one point in A that is also in B.

3.2.3 Equals

A equals B if all points in the two objects are in common.

3.2.4 Covered by or inside or contained in
A is contained in B if all points in A are in B. There may be points in B that are not in A.

3.2.5 Covers or contains
A contains B iff B is contained in A.

Based on the placement of the objects in the space, relationships with respect to direction such as north, south, east, west and so on. These relationship is not as easy to identify irregular shape and overlap of spatial objects. Generally the Euclidean and Manhattan measures are used to compute the distance between two points.

3.3 Spatial Rules
Spatial rules can be generated to depict the relationship between structure of spatial objects. The fundamental rules of spatial data are :

3.3.1 Spatial characteristic rules
It describe the data. For eg. In Rajasthan the average income is Rs. 20,000.

3.3.2 Spatial discriminate rules
It describe the differences between different classes of the data. They describe the features that differentiate the different classes. For eg. In Rajasthan the average income is Rs. 20,000 while in NCR the average income is Rs. 30,000.

3.3.3 Characterization
It is the process of finding a description for a dataset or some subset thereof.

3.3.4 Trend detection
It viewed as a continuous variation in single or more nonspatial attribute values for spatial objects as move away from another spatial object. One of the trend detection techniques is krigging to predict the location from outside the sample. For e.g. The average income may decrease as the proximity to the rural increases.

3.3.5 Spatial Association rules
These are implications of single set of data by other. For e.g. in Rajasthan the average income for person living near rural is Rs. 15,000. It discovers uncovering relationship from spatially related dataset and used to describe patterns of the database. It is used to find the occurrence of an event Y in the neighbourhoods of another event X in spatiotemporal data [28]. At least one of the antecedent or the consequent should contain some spatial predicate in the rule as follows.

3.3.5.1 Non geo-referenced antecedent and geo-referenced consequent:
All basic crops are sailed near to apartments.

3.3.5.2 Geo-referenced antecedent and non geo-referenced consequent:
If an apartment is cited in main city, it is costly.

3.3.5.3 Geo-referenced antecedent and geo-referenced consequent:
Any apartment which is close to down town is city. For spatial association rules, support and confidence are defined identically to that for regular association rules. Spatial Association is used to find positive and negative association which extracts multilevel interesting patterns in spatial and nonspatial predicates using topological relation [23]. A multilevel Association Rule has been generated to find association between the data in a large database. [29] explain some patterns that are in spatial time series data.

3.3.6 Spatial Classification
These are used to partition sets of spatial objects. Spatial objects could be classified using nonspatial attributes, spatial predicated, or spatial and nonspatial attributes. The spatial object has been classified by using its attributes. Each classified object is assigned a class. It is the method of finding a set of rules to decide the class of spatial object [30]. It classify attributes of the object along with neighboring objects with their spatial relation. The spatial classification methods such as decision trees (C4.5), artificial neural network, remote sensing, spatial autoregressive regression are used to find the group of the spatial objects together. Concept hierarchies, sampling, generalization and progressive refinement techniques can be used to improve efficiency. The classification problem is applied in the area of transporting for dividing spatial locations based on the area.

3.3.7 Spatial Clustering
It is used by grouping to discover similarity in between spatial dataset related to features found in the actual spatial data. Set of like elements are grouped into one clusters and elements from different clusters are not alike. Spatial clustering is based on the distance and direction relation. Spatial clustering techniques used ranges from partition, hierarchical, density based and grid based method.

Table 3 summarizes SDM techniques used in various domain as listed from various literature.

3.4 Spatial Data Mining Tools
DBMiner (DBlearn), GeoMiner are open source tools, used for data mining and query language, geomining and query language respectively and developed by Data Mining Research Group, Simon Fraser University, Canada. GeoDA(ESDA,STARS) is also open source Python language based tool which supports spatial autocorrelation statistics, spatial regression. Another open source Java language based tool is Weka-GDPM, developed by University of Waikato, NZ, which supports several standard data mining tasks. R language (sp, rgdal, rgeos) is also open source tool, compatible with C, FORTRAN, and Python language and support to analysis statistical and graphical techniques. Descrates is open source tool for Python language to visualize and analysis source data and display the outcome of classification on the plot. ArcGIS (ArcView, ArcInfo, ArcEditor) is paid tool, developed by Environmental Systems Research Institute (ESRI) used for Web API, Python, .NET language and used for spatial analysis and modeling features. It includes surface, network analysis, overlay, interpolation analysis and geo-statistical modeling techniques.

4. ISSUES, CHALLENGES & SOLUTION ARCHITECTURE
With the explosive growth in GIS data is viewing in respect of data integration and mining huge spatial data volume. To overcome this issue, architecture is proposed to deal the problems of huge volume spatial data integration on the basis of analysis of spatial data from literature is described in Fig.4. Data warehouse technique is providing the facility for data integration and deposits summarized data. The fast growing data is the origin of big data. In recent the big data approach has coined for mining spatial data using parallel algorithm on Hadoop and MapReduce architecture in a distributed environment to deal with such huge amount of datasets. We proposed a MRPrePost, hadoop architecture based, Pre-Post

algorithm, used for mining big data in this paper. MRPrePost is hybrid of the Dis-Eclat [31] and used as association rule mining to find frequent item set from huge spatial data sets.

Table 3 Spatial Data Mining techniques used

SDM Domain	Usage	Technique/Method	Reference
Forest	Identify false alarms in forest fire hotspots	Region growing method, Hough transform	Satellite images are used for experiments [3]
	GIS based fireproof system	Frequency theory based method	Sample spatial dataset [4]
	to evaluate forest extent changes	Spatial Data Mining, Back propagation algorithm	Satellite images are taken as dataset [5]
Transport	To increase the effectiveness of railway MIS such as monitoring, railway tracks, and geographical spread	Association rule mining, classification, forecast, trend analysis and planning	[6]
	Represent link between the GIS street data and roadway connections	Object oriented modeling transportation	[7]
Warehouse	Improving spatial data mining effectiveness	Spatial data cube	Spatial data from warehouse [8]
Agriculture	Precision agriculture	Cross-validation technique	Spatial dataset [9]
	Crop yield prediction for wheat	Neural Network	Experiments are made on satellite images [10]
	accessing the quality of soil, management of water resources.	GIS and Spatial Data mining	[11]
Tourism Management	integrating the ICT techniques with tourism	association technique	[12]
Land allocation	provide the selection a model for land use, illegal land fills	Location prediction	[13]
Meteorology	Estimation of rainfall for homogenous monsoon region	genetic approach and correlation	[14]
	forecast the rainfall of Rajasthan state	Multiple liner regression	[15]
	Regionwise rainfall fluctuation	classification	[16]
	Estimation of rainfall	Spatial interpolation and association rule	[17]

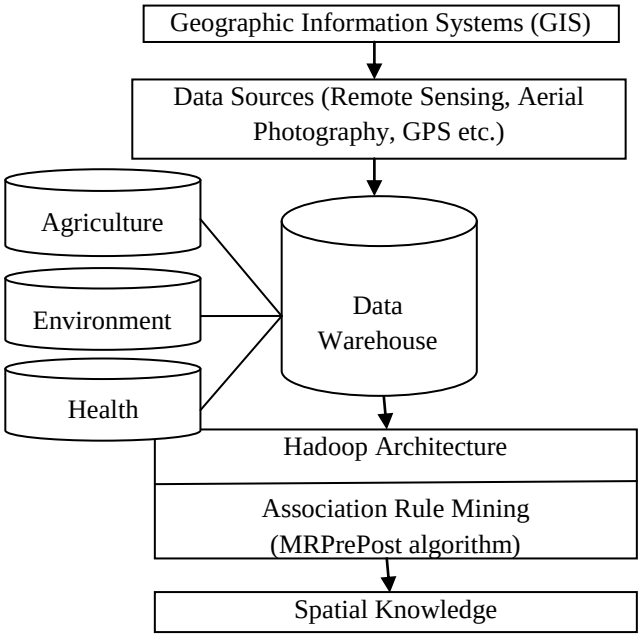

Fig.4. Big Data Approach – Integration of GIS Data

5. CONCLUSION

This paper provides with a detailed knowledge on spatial data and its characteristics. A detailed analysis and description of GIS data sources, data formats and data representation is presented from various literatures. The data mining algorithms used for different applications in GIS are also detailed. The other major challenge of GIS databases can be viewed as volume and data formats in this work we have proposed architecture for data integration using data warehouse approach in GIS. The key challenge in GIS is integration of large volume of data. To deal this challenge we have viewed the problem as bigdata and represented the same in our proposed architecture design. In prospect the proposed solution would be implemented and tested for agriculture domain.

6. REFERENCES

[1] Maguire, D. J., Goodchild, M. F., and Rhind, D. W. 1991. Geographical Information Systems, pp.9-20

[2] Heywood, I., Cornelius, S., Carver, S., Raju, and S. An Introduction to Geographical Information Systems, Second Edition, Pearson Education, pp.10.

[3] Sengchuan, T. 2003. Spatial Data Mining: Clustering of Hot Spots and Pattern Recognition. IEEE. pp.3685-3687

[4] Liang,Y., and Fuling, B. 2007. An Incremental Data Mining Method for Spatial Association Rule in GIS Based Fireproof System. IEEE. pp.5983-5986.

[5] Jayasinghe, P.K.S.C., and Masao, Y. 2013. Spatial data mining technique to evaluate forest extent changes using GIS and Remote Sensing.

[6] Wei, X., Yong, Q., Houkuan, and H. 2003. The Application of Spatial Data Mining in Railway Geographic Information Systems. IEEE. pp.1467-1471

[7] Marzolf, F., Trépanier, M., and Langevin. 2006. A Road network monitoring algorithms and a case study. Journal of Computer and Operation Research, pp.3494–3507.

[8] Yuanzhi, Z., XieKunqing, M., Xiujun, X., Dan, C., and Tang S. 2005. Spatial Data Cube: Provides Better Support for Spatial Data Mining. IEEE. pp.795-798

[9] Rub, G., and Brenning, A. 2010. Data Mining in Precision Agriculture: Management of Spatial Information, Computational Intelligence for Knowledge Based System Design, Volume 6178, pp. 350-359.

[10] Stathakis, D., Savin, I., and Nègre T. Neuro-Fuzzy Modeling for Crop Yield Prediction, The International Archives of the Photogrammetry, Remote Sensing and Spatial Info. Sc., Vol. 34, pp.1-4.

[11] Vaagh, Y. 2012. The application of a visual data mining framework to determine soil, climate and land use relationships. Journal of Procedia Eng. 32, pp.299–306 .

[12] Buhalis, D., and Law, R. 2008. Progress in information technology and tourism management,The state of eTourism research. Journal of Tourism Mgmt.

[13] Chakraaborty, A., Mandal, J.K., Chandrabanshi, S.B., and Sarkaar, S. 2013. A GIS Anchored system for selection of utility service stations through Hierarchical Clustering. International Conference on Computational Intelligence: Modeling techniques and Application, CIMTA

[14] Kashid, S.S., and Maity, R., 2012. Prediction of monthly rainfall on homogenous monsoon regions of India based on large scale circulation patterns using Genetic Programming. Journal of Hydrology, pp.26-41.

[15] Vyas, P., 2015. To predict rainfall in desert area of Rajasthan using data mining techniques. vol.3, no.5.

[16] Priya, R.L., and Manimannan, G., 2014. Rainfall fluctuation and regionwise classificatrion in Tamilnadu using geographical information system. IOSR Journal of Mathematics (IOSR-JM), vol. 10, pp.5-12.

[17] Teegavarapu, R. S. V., 2009. Estimation of missing precipitation records integrating surface interpolation techniques and spatio-temporal association rules. Journal of Hydroinformatics, vol. 11, no. 2, pp.133–146.

[18] http://www.newdesignfile.com/postpic/2013/04/vector-and-raster-data-gis_132173.JPG

[19] http://www.vermessungsseiten.de/gis/vector_raster.gif

[20] Niebles, J.C., Wang, H., and Fei-Fei, L. 2008 Unsupervised learning of human action categories using spatial–temporal words. Int. J. Compute. Vis.79(3), pp.299–318.

[21] Shekhar, S., Zhang, P., Huang, Y., and Vatsavai, R.R., (2003): Trends in spatial data mining.

[22] Lee, A.J.T., Hong, R.W., Ko, W.M., Tsao, W.K., and Lin, H.H. 2007. Mining spatial association rules in image databases. Inform. Sci. 177, pp.1593–1608.

[23] Du, S., Qin, Q., Wang, Q., and Ma, H. 2008. Evaluating structural and topological consistency of complex regions with broad boundaries in multi-resolution spatial databases. Information Sci. 178, pp.52–68.

[24] Pop III, A., Burnett, R.T., Thurston, M.J., and Calle, E.E., and Krewski. 2004. Cardiovascular Mortality and Long-Term Exposure to Particulate Air Pollution. Circulation 109, pp.71–77.

[25] Egenhofer, M. 1994. Spatial SQL A Query and Presentation Language. IEEE Transactions and Data Engineering 6, pp.86–95.

[26] Spatial Data Mining, Winter School on 'Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets, pp. 153-166.

[27] Dunham, M.H. 2006. Basic Data Mining Tasks. Singapore, Pearson Education.

[28] Mennis, J., and Liu, J. W. 2005. Mining Association Rules in SpatioTemporal Data: An Analysis of Urban Socioeconomic and Land Cover Change. Transactions in GIS, 9(1), pp.5-17.

[29] Shekhar, S., Evans, M. R., Kang, J. M., and Mohan, P. 2011. Identifying Patterns in Spatial Information: A Survey of Methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(3), pp.193-214.

[30] Brinkoff, T., and Kriegel, H.P., 1994. The Impact of Global Clustering on Spatial Database Systems. Proceedings of the 2Uth VLDB Conference, pp.168–179.

[31] Moens S., Aksehirli E., and Goethals B. 2013. Frequent Itemset Mining for Big Data, IEEE Int. Conf. on Big Data, IEEE, pp.111-118.