

Edith Cowan University
Research Online

ECU Publications Post 2013

1-1-2014

Geospatial data pre-processing on watershed datasets: A GIS approach

Sreedhar Nallan

Edith Cowan University, acharyasree@gmail.com

Leisa Armstrong


Edith Cowan University, l.armstrong@ecu.edu.au

Barry Croke

Amiya K. Tripathy

Edith Cowan University, a.tripathy@ecu.edu.au

Follow this and additional works at: <https://ro.ecu.edu.au/ecuworkspost2013>

 Part of the [Agricultural Science Commons](#), [Databases and Information Systems Commons](#), and the [Hydrology Commons](#)

Nallan, S. , Armstrong, L. , Croke, B., & Tripathy, A. K. (2014). Geospatial data pre-processing on watershed datasets: A GIS approach. *Proceedings of Asian Federation for Information Technology in Agriculture*. (pp. 328-336). Perth, W.A. Australian Society of Information and Communication Technologies in Agriculture. Available [here](#)
This Conference Proceeding is posted at Research Online.
<https://ro.ecu.edu.au/ecuworkspost2013/851>

Geospatial data pre-processing on watershed datasets: A GIS approach

Sreedhar Nallan¹, Leisa Armstrong¹, Barry Croke², Amiya Kumar Tripathy³

¹ *School of Computer and Security Science, Edith Cowan University, Perth, Australia*

² *Integrated Catchment Assessment and Management (iCAM) Centre and the National Centre for Groundwater Research and Training, The Fenner School of Environment and Society, The Australian National University, Canberra, Australia*

³ *Department of Computer Engineering, Don Bosco Institute of Technology (University of Mumbai), Mumbai, India*

Email: snallan@our.ecu.edu.au

Abstract

Spatial data mining helps to identify interesting patterns from the spatial data sets. However, geo spatial data requires substantial data pre-processing before data can be interrogated further using data mining techniques. Multi-dimensional spatial data has been used to explain the spatial analysis and SOLAP for pre-processing data. This paper examines some of the methods for pre-processing of the data using Arc GIS 10.2 and Spatial Analyst with a case study dataset of a watershed.

Key Words: geospatial, data mining, pre-processing, watershed

Introduction

Spatial data mining helps to identify unknown patterns in large databases of spatial data. It helps the researchers and policy makers to understand the spatial and temporal variability of the information through visualisation of the data. Unlike the classical data mining, knowledge discovery from the spatial data with multi-dimensional parameters is a tedious effort (Guo and Mennis 2009). It is equally time consuming task to perform data pre-processing before using the spatial data for data mining (Ester, Frommelt et al. 2000, Sharma 2006). This paper examines the spatial datasets and suggests some of the methods which can be used for pre-processing data using an example from watershed dataset. The methods and properties mentioned in this paper are relevant to Arc Map 10.2 software with Spatial Analyst extension (ESRI 2013).

Data mining

The basic definition of data mining is a ‘process of discovering non trivial, interesting and unknown patterns from the databases’ (Fayyad, Piatetsky-Shapiro et al. 2010). From a managerial perspective, data mining can assist in decision making by providing the means to find hidden patterns and trends in data that is not immediately apparent from summarizing the data. The major steps involved in the data mining process include a) defining the problem, b) obtaining background knowledge, c) appropriate data selection, d) pre-processing data to fill the gaps and removing unwanted or erroneous data, e) data mining and f) results evaluation (Marvin and John 2003). Pre-processing of data is a time consuming exercise in formatting multi format data into unified format compatible to be used in any data mining software.

Geospatial data mining

Unlike classical data, spatial data consists of the interrelated data at a spatial scale. The mining of spatial data helps to understand the inter-relation between different spatial entities which are depended on other and effecting the changes at one location to other (Mukhlash and Sitohang 2012). The data mining process is complex in spatial data as compared to other non-spatial relation data (Ester, Kriegel et al. 2001). For example, Figure 1 shows the location of water harvesting structures (check dams) in different locations of a watershed having different soil types. The impact of the location of water harvesting nearer and far in different soils will affect the ground water table. Spatial data mining interrogation helps to understand the impact such development in a better manner.

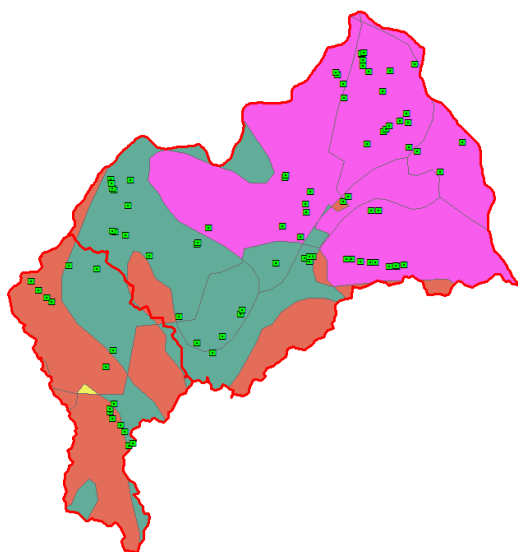


Figure 1: Locations of Water Harvesting Structures in different Soils

Spatial datasets

Spatial datasets consists of vector and raster layers. The vector layers will have different thematic layers mainly classified as a) point layers, b) polygon layers or c) line layers (Fotheringham and Rogerson 2005). For example, the point layers represent the locations of wells in a village. These are represented with an exact location of a ground water well using latitude and longitude values. The polygon layer consists of an irregular or regular shape with defined boundaries such as a watershed boundary. The line layer represents the roads, river network or the flow direction of the water. The raster layers are mostly from the Satellite Imagery having a pixel based grid values. Each pixel represents certain area on the ground based on its resolution. The grid layers such as digital elevation model (DEM) obtained from ASTER 30M satellite dataset is used to derive the stream network in a watershed. All these layers are inter-related to one selected area with multi-dimensional data formats with the data availability at different time scales and different space scales. The procedures to integrate these data sets and utilize it for data mining are discussed in the following sections.

Pre-processing of the data depend on the type and the quality of the selected datasets (Han, Kamber et al. 2012). Spatial datasets pre-processing is time consuming and higher effort task (Bogorny, Engel et al. 2006). The spatial data pre-processing is carried out with the steps including a) reduce the non-relevant data which has no spatial attributes (or not relevant to the study) b) perform transformation of the data using instances of granularity and featured

granularity (Camossi, Bertolotto et al. 2003) and c) format into single flat file. Pre-processing largest spatial data sets are being carried out utilizing Bayesian networks, principle component analysis (PCA), non-negative matrix factorization (NMF) and k-means clustering (Hyvönen, Junttila et al. 2007). The multi-dimensional spatial data can be formatted to unified format by generating actual data into classes (Chen, Lin et al. 2011).

The pre-processing pays more attention on incomplete data, inaccurate data, repetitive data and inconsistent data. While many researchers have attempted to explore the ways of data pre-processing, non-spatial data or data with sample data sets, still little research has been reported on the techniques used for pre-processing spatial datasets (Wang, Li et al. 2003, Yanli, Ramanathan et al. 2011). This paper discusses how pre-processing methods can be used to solve some of the issues related to geospatial data on a) the handling the missing values in a time series data and b) the scaling the point values to polygons.

Pre-processing Watershed Spatial Dataset

Geo spatial data plays an important role in watershed related studies. The data on watersheds is multi-dimensional and consists of different formats both spatially and temporally. It's a time consuming task to integrate the multi-dimensional data sets into common format for proper data mining. This section explores the techniques to be used for pre-processing the spatial data using ArcMap 10.2 with Spatial Analyst Extension. Spatial data on soil, geology, ground water, rainfall and water harvesting structures plays vital role in the development process of watershed. The daily rainfall data and the watershed boundaries data have been utilized in exploring the pre-processing techniques in this paper.

The initial task in spatial pre-processing looks at detecting the outliers, making uniform projections and scaling the multi-level data into single format. The topological relations characterize the relation between two geographic features in a spatial data. With the spatial functions available in spatial analyst such as *intersect* or *union*, the data which is not relevant to the study area can be eliminated.

Spatial Outliers in Point data

The point data such as rainfall at particular location, or water level in a bore well is represented as spatial data using the geographical positioning values such as Latitude and Longitude. Initial pre-processing for this data goes with mapping the data onto the GIS map and identifying the outliers which are out of the selected area (Fig. 2). The outliers are defined as an extreme value not relevant to the current subject. This can be due to a) erroneous representation of the coordinates or b) non relevant locations out of the study area.

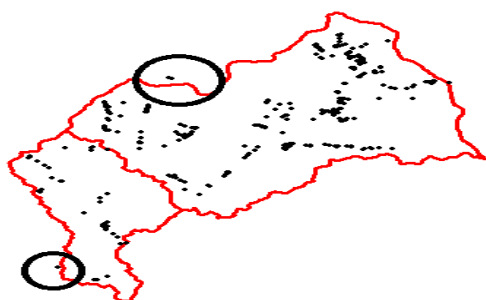


Figure 2: Points showing outside of the study of interest

Spatial outliers in Polygon data

Polygon data is used to represent certain geographic area as a thematic layer. This can be an administrative boundary, hydrological boundary or a soil type (Fig. 3). While dealing with multi polygonal datasets, pre-processing requires a great attention and an understanding of the inter connectivity between the different polygons is essential. The pre-processing for a unifying format can then be undertaken following this initial analysis of the data

Pre-processing for unified format

Using the GIS software methods available in as *Clip*, *nion*, merge methods the different polygons can be made into unified format. Spatial data consists of the relation between the objects with neighbourhood relations. Pre-processing this data can be made utilizing ‘*Near*’ method to assign the nearest rain gauge value to the village data when the rain gauges are less and villages are more. In other case, the mean value of the rain gauges rainfall can be assigned to the village as an average rainfall.

Depending on the source data, the spatial boundaries can be deviated. This can be due to different projections or different resolution images used as base data. While working with multiple data sets of such source, there will be ambiguity to deal with the boundaries. Fig 3-a represents a watershed boundary derived from ASTER 30M DEM. While this is compared with other demographic information such as village level population statistics, this need to be re organized to match with the villages covered in each watershed (Fig 3-b). The ambiguity can also be due to sub watersheds and a bigger watershed, as the derivations of the watersheds can be delineated from different sources of imagery with different resolutions (Fig 3-c). Other important fact deal with watershed is soil properties. Soil properties do not match with the administrative boundary (village) or hydrological boundary (watershed). For an effect assessment of watersheds, the relevant soil properties need to be assigned to the watershed properties for understanding the water retention properties in that watershed (Fig 3-d).

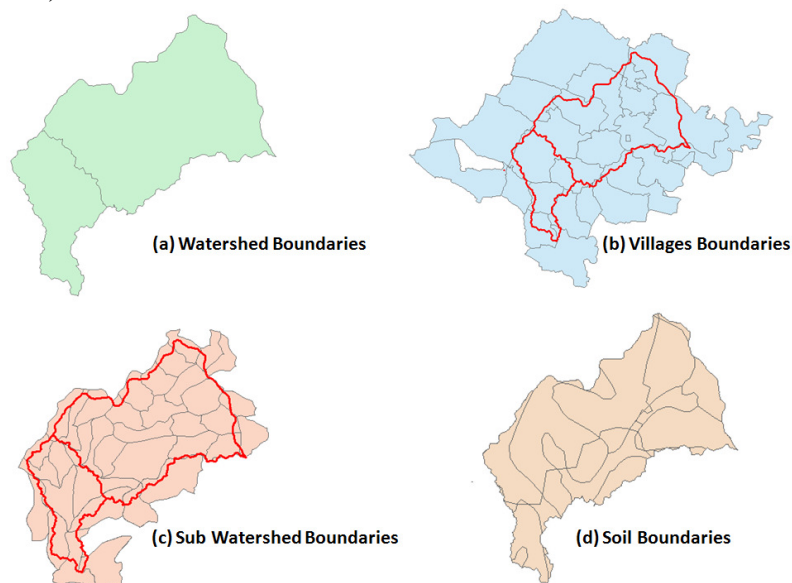


Figure 3. Polygon shape of different parameters for a given area

Missing values

Most of the spatio-temporal data in huge databases have inconsistent data or inaccurate data or incomplete data. Incomplete data refers to the data which is not available for all parameters or available for limited period. Inaccurate data goes with wrong entry of the values or the values out of the range. Inconsistent data goes with different units of the data used, and different codes given for the same dataset in different places.

While the time series data is a continuous data, there is always a big challenge to deal with the missing values in a given time series data. This can be of two types, spatial or temporal. Temporal missing of data can be attributed as the values for a particular period for every interval is not available (Wang, Shi et al. 2005). Spatial missing of data can be due to non-availability of data for some locations. Spatial Online Analytical Processing (SOLAP) techniques can be used to identify the gaps (Fig 3). In both the cases, the data need to be pre-processed with the actual factual information if available or using spatial analyst extension tool set. The missing data can be filled with using different interpolation techniques such as Spline, Krig or Inverse Distance Weightage. For example, rainfall data for a selected site for the period 1989-1994 is given in the Figure 3. The missing values of the years 1990 and 1993 has been filled using Spatial Analyst – IDW interpolation technique. The values of the missing data can be extracted from the Raster interpolated layer and can be added to the original dataset (Fig. 4, 5).

Station	Latitude	Longitude	Year	Annual Rainfal
Station A	15.1303	77.6778	1989	654
Station B	15.114	77.4651	1989	705
Station D	15.2391	77.8248	1989	584
Station C	15.3337	77.5973	1989	557
Station A	15.1303	77.6778	1990	560
Station C	15.3337	77.5973	1990	728
Station A	15.1303	77.6778	1991	545
Station B	15.114	77.4651	1991	327
Station D	15.2391	77.8248	1991	483
Station C	15.3337	77.5973	1991	606
Station A	15.1303	77.6778	1992	548
Station B	15.114	77.4651	1992	460
Station D	15.2391	77.8248	1992	592
Station C	15.3337	77.5973	1992	607
Station A	15.1303	77.6778	1993	511
Station D	15.2391	77.8248	1993	785
Station C	15.3337	77.5973	1993	546
Station A	15.1303	77.6778	1994	526
Station B	15.114	77.4651	1994	346
Station D	15.2391	77.8248	1994	550
Station C	15.3337	77.5973	1994	261

Station	1989	1990	1991	1992	1993	1994
Station A	654	560	545	548	511	526
Station B	705		327	460		346
Station C	557	728	606	607	546	261
Station D	584		483	592	785	550

Figure 4: Annual Rainfall Year wise(left) – Formatted data into Station wise and Year wise Rainfall to identify missing values (right)

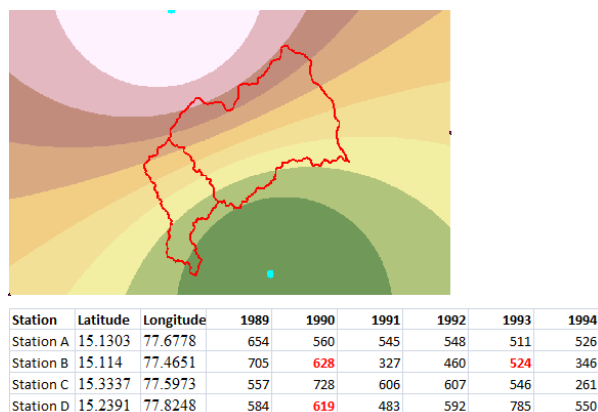


Figure 5: Missing Values filled with Interpolated Values using IDW method of Spatial Analyst

Projection

Features for different entities collected from different sources will have scale issue. The different scales of the multi-source data makes unfeasible to make a unified format (Weibel and Dutton 1999). Pre-processing of spatial data needs to be addressed this issue while preparing the data for spatial data mining. Due to different projections of the spatial datasets, all datasets need to be made into unique projections before being used for analysis. This is important especially while working at the micro scale where small differences can also affect the results. Using the 'Project' method of Spatial Analyst tool, the spatial data can be assigned with a unique Projection.

Conversion of Spatial value at different granularities (Point to Polygon)

The scale of the dataset is another important issue to be addressed when collating different spatial data sets. The data collected at point level refers to a location Whereas, if some other data is available at polygon level, this point data need to be converted into polygonal data, to make it into unified format. Often, this issue can be overcome by using spatial interpolation techniques (Atkinson and Tate 2000). For example, for a given study area, rainfall data is available for 4 rain gauge s as daily data. For example. if village level annual rainfall is needed to be calculated, using these points, the daily rainfall can be initially aggregated into annual rainfall. The annual rainfall can be interpolated for required year to create a spatial grid. Finally, using Spatial Analyst function of 'Zonal statistics as Table', the mean value from each spatial raster for each village polygon can be extracted (Fig. 6).

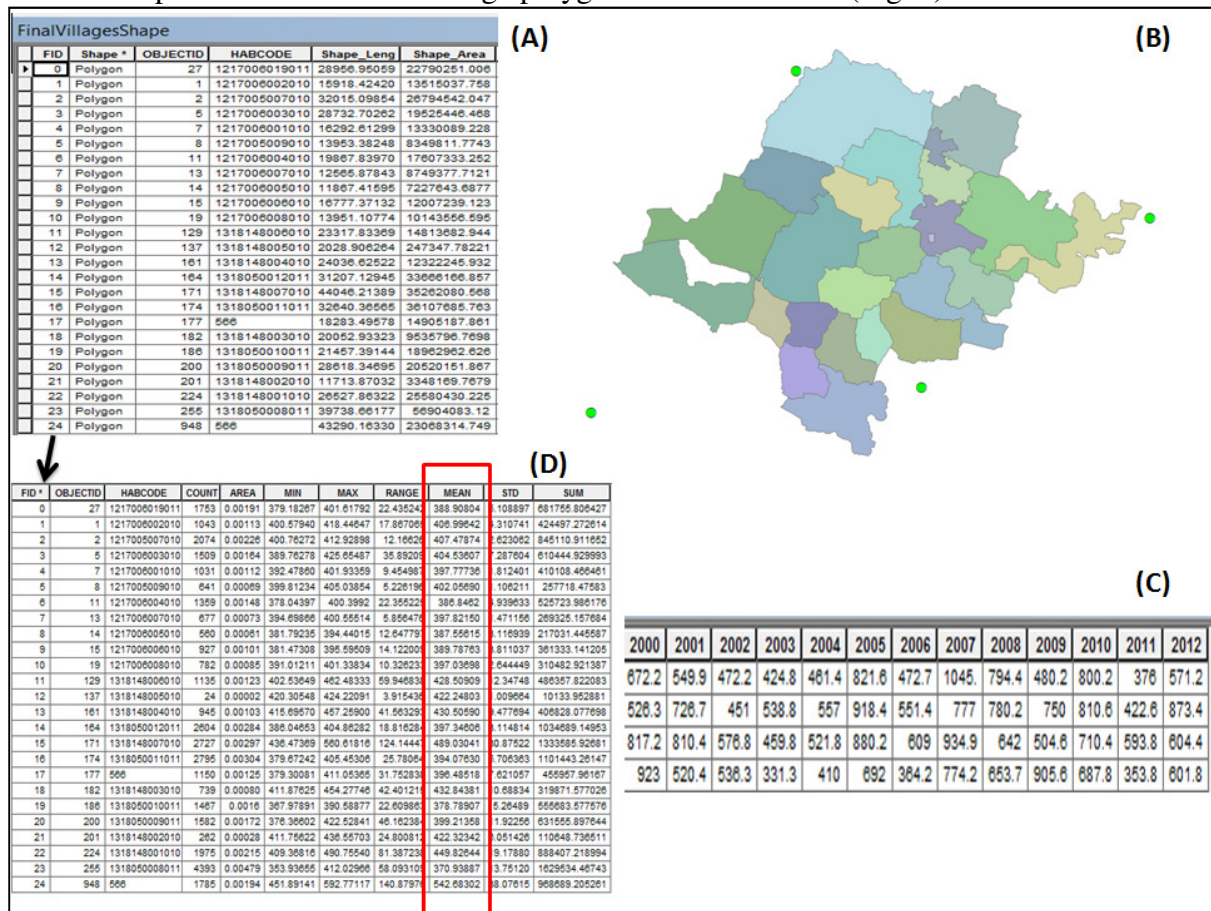


Figure 6: (A) Spatial data with Villages (B) Locations of Rain Gauges (as dots) and Villages Shape (C) Year wise data for each Rain Gauge (D) Extracted values from Zonal Statistics of each Village

Spatial granularity dealing with multiple formats

Geospatial databases often represent the same real world object in different formats (Volz 2006). These formats can be of two types; 1) spatial data for different entities represented as per its boundaries 2) spatial data with different structure. For example, a watershed dataset consists of different formatted data both spatial and non-spatial. For example soil and geology data can be differentiated in shape files with varying attributes. While utilizing both these datasets, for spatial data mining, a third spatial dataset need to be created with all the attributes of both the datasets. In the second case, a rainfall data is represented at level for each year cannot be analysed with the spatial analysis, unless it is formatted into as rows with its location coordinates and yearly rainfall in columns (Table 2).

Table 1: Annual Rainfall a) Vertical b) Spatial Pattern

a) Vertical Pattern with Station wise / Year wise Rainfall

Station	Year	Latitude	Longitude	Rainfall (mm)
A	2008	15.13032	77.67776	794.4
A	2009	15.13032	77.67776	480.2
A	2010	15.13032	77.67776	800.2
A	2011	15.13032	77.67776	376
A	2012	15.13032	77.67776	571.2
B	2008	15.11398	77.4651	780.2
B	2009	15.11398	77.4651	750
B	2010	15.11398	77.4651	810.6
B	2011	15.11398	77.4651	422.6
B	2012	15.11398	77.4651	873.4

Station	Year	Latitude	Longitude	Rainfall (mm)
C	2008	15.23914	77.82478	642
C	2009	15.23914	77.82478	504.6
C	2010	15.23914	77.82478	710.4
C	2011	15.23914	77.82478	593.8
C	2012	15.23914	77.82478	604.4
D	2008	15.33372	77.59727	653.7
D	2009	15.33372	77.59727	905.6
D	2010	15.33372	77.59727	687.8
D	2011	15.33372	77.59727	353.8
D	2012	15.33372	77.59727	601.8

b) Dataset mapped to GIS software for spatial analysis derived from (a) above

Station	Latitude	Longitude	2008	2009	2010	2011	2012
A	15.13032	77.67776	794.4	480.2	800.2	376	571.2
B	15.11398	77.4651	780.2	750	810.6	422.6	873.4
C	15.23914	77.82478	642	504.6	710.4	593.8	604.4
D	15.33372	77.59727	653.7	905.6	687.8	353.8	601.8

Conclusion and Future work

Spatial data pre-processing is time consuming component of any spatial data mining process. Statistical approaches have also been used by researchers to fill the gaps in the data (Hasan and Croke 2013). This paper reports on the issues related to pre-processing of this data by utilizing the different GIS and database techniques in a holistic manner, with a view to make this process less time-consuming and more a efficient task. This paper discussed some of the pre-processing methods on spatial datasets to remove outliers, filling missing data and dealing with different formats of data, projection issues, converting point data into polygonal and dealing with multiple formats data.

These proposed pre-processing techniques can be used for geospatial data mining of topology and climatic studies for Apart from inverse distance weightage, Spline and Krig techniques has been used for spatial interpolation of the rainfall data. This paper examined the

application of some of the Spatial Analyst methods available in Arc Map 10.2 which can be used in pre-processing the spatial datasets. Further research on validation of the output using different techniques in different conditions can help to understand the utilization of such methods in a robust manner.

Acknowledgement

This research is supported by Edith Cowan University International Post Graduate Research Students scholarship.

References

- Atkinson, P. M. and N. J. Tate (2000). "Spatial Scale Problems and Geostatistical Solutions: A Review." *The Professional Geographer* 52(4): 607-623.
- Bogorny, V., et al. (2006). GEOARM: an Interoperable Framework to Improve Geographic Data Preprocessing and Spatial Association Rule Mining. SEKE.
- Camossi, E., et al. (2003). Issues on modeling spatial granularities. COSIT.
- Chen, J., et al. (2011). Extracting spatial association rules from the maximum frequent itemsets based on Boolean matrix. *Geoinformatics, 2011 19th International Conference on, IEEE*.
- ESRI (2013). Arc Map 10.2.
- Ester, M., et al. (2000). "Spatial data mining: database primitives, algorithms and efficient DBMS Support." *Data Mining and Knowledge Discovery* 4(2): 193-216.
- Ester, M., et al. (2001). Algorithms and applications for spatial data mining. *Geographic Data Mining and Knowledge Discovery, Taylor and Francis*.
- Fayyad, U., et al. (2010). "Knowledge discovery and data mining: towards a unifying framework. 1996." *KDD Proceedings, AAAI*.
- Fotheringham, S. and P. Rogerson (2005). *Spatial analysis and GIS, CRC Press*.
- Guo, D. and J. Mennis (2009). "Spatial data mining and geographic knowledge discovery - An introduction." *Computers, Environment and Urban Systems* 33: 6.
- Han, J., et al. (2012). *Data Mining: Concepts and Techniques 3rd Edition*. Gurgaon, India, Elsevier India Private Limited.
- Hasan, M. M. and B. F. W. Croke (2013). Filling gaps in daily rainfall data: a statistical approach. *20th International Congress on Modelling and Simulation, Adelaide, Australia*.
- Hyvönen, S., et al. (2007). Pre-processing Large Spatial Data Sets with Bayesian Methods. *Knowledge Discovery in Databases: PKDD 2007*. J. Kok, J. Koronacki, R. Lopez de Mantaras et al., Springer Berlin Heidelberg. 4702: 498-505.
- Marvin, L. B. and F. K. John (2003). "Data mining and the impact of missing data." *Industrial Management & Data Systems* 103(8): 611-621.
- Mukhlash, I. and B. Sitohang (2012). "Spatial data preprocessing for mining spatial association rule with conventional association mining algorithms."
- Sharma, A. (2006). *Spatial data mining for drought monitoring: An approach using temporal NDVI and rainfall relationship. Geo-information Science and Earth Observation. The Netherlands, International Institute for Geo-information Science and Earth Observation. Masters Thesis: 87.*

Volz, S. (2006). An iterative approach for matching multiple representations of street data. ISPRS Workshop, Multiple representation and interoperability of spatial data, Hanover, Germany.

Wang, S., et al. (2003). "Cloud Model-Based Spatial Data Mining." *Annals of GIS* 9(1): 60-70.

Wang, S., et al. (2005). Attribute Uncertainty in GIS Data. *Fuzzy Systems and Knowledge Discovery*. L. Wang and Y. Jin, Springer Berlin Heidelberg, 3614: 614-623.

Weibel, R. and G. Dutton (1999). "Generalising spatial data and dealing with multiple representations." *Geographical information systems* 1: 125-155.

Yanli, Z., et al. (2011). "Web-based spatial decision support system and watershed management with a case study." *International Journal of Geosciences* 2(3): 195-203.
