Journal of the Earth and Space Physics, Vol. 42, No. 4, Winter 2017, PP. 35-41

Evaluation of groundwater vulnerability using data mining technique in Hashtgerd plain

Javadi, S.[1*] and Hashemy, M.[1]

1. Assistant Professor, Department of Water Engineering, College of Abouraihan, University of Tehran, Iran

(Received: 18 Oct 2015, Accepted: 14 Jun 2016)

Abstract

Groundwater vulnerability assessment is an effective informative method to provide basis for determining source of pollution. Vulnerability maps are employed as an important solution in order to handle entrance of pollution into the aquifers. A common way to develop groundwater vulnerability map is DRASTIC index. Meanwhile, application of the method is not easy for any aquifer due to choosing appropriate constant values of weights and ranks. Clustering technique would be an influential method for regionalization of groundwater flow zone to facilitate vulnerability assessment of groundwater aquifers. In this study, a new approach using k-means clustering is applied to make vulnerability maps. Four features of depth to groundwater, hydraulic conductivity, recharge value and vadose zone are considered at the same time as features of clustering. Five regions are recognized out of the Hashtgerd plain. Each zone corresponds to a different level of vulnerability. The results show that clustering provides a more realistic vulnerability map so that, Pearson's correlation coefficients between nitrate concentrations and clustering vulnerability is 72%.

Keywords: Groundwater, Vulnerability assessment, Clustering, Data mining.

1. Introduction

Due to simple operation and no needs for expensive infrastructure construction to convey water from a source to farm lands, groundwater becomes the most important sources of agricultural water supply in Iran.

However, the contamination of aquifers is a major concern in many countries, specifically in areas without effective groundwater protection and management. Therefore, groundwater vulnerability assessment can be one of the effective informative methods to provide a basis for determining source of pollution. Assessment of groundwater vulnerability is often done by intrinsic vulnerability, which considers hydro-geological conditions. The concept of vulnerability of aquifers was introduced for the first time by Marget in 1986 (Margat, 1968). The first definition of vulnerability was proposed by Marget and it means the degree of groundwater contamination by pollution reaching the groundwater system (Margat, 1968). Overlay and index method could be mentioned as existing method to assess intrinsic vulnerability of groundwater (Wang et al., 2012). Moreover the vulnerability index is relatively, dimensionless and immeasurable and depends on the hydrogeology and geology characteristics of the aquifer. Since then, many researchers applied many methods and techniques to provide a standard way for evaluation of vulnerability. Many approaches and techniques have been utilized to evaluate vulnerability, for instance some vulnerability indexes are GOD, DRASTIC (Aller et al., 1987), AVI rating system, SEEPAGE, SINTACS, ISIS, EPIK, and DIVERSITY. As it was mentioned before, it should be noted that all of methods are relative and dimensionless, using various data depending on the sort of aquifer.

Compared to other models, DRASTIC model, is an overlay and index method, that is the most popular index that have been used by many researches. The DRASTIC model is easy to implement and provides a good basis for assessment of groundwater vulnerability in facing contamination (Baalousha, 2006). Also it needs a relatively small amount of data that is often available for many aquifers (Wang et al., 2012).

Although, the DRASTIC method is

[*]Corresponding author: E-mail: javadis@ut.ac.ir

mostly accepted by the researcher, but it has some disadvantages and shortcomings. According to different experiences out of the method application, the subjectivity in assigning numerical values to the descriptive entities is introduced as a limitation of applying DRASTIC (Javadi et al., 2011a). Furthermore, employing relative ranks and weights for different attributes without considering the position and place are other drawbacks. Therefore, the same weights and rating values are used everywhere since the influences of regional characteristics that are not considered in the method (Javadi et al., 2011a, b). However, there have been numerous study carried out to test and modify validity of DRASTIC based algorithms (Panagopoulos et al., 2006; Neshat et al., 2014; Niknam et al., 2009; Nobre et al., 2007; Saidi et al., 2011; Javadi et al., 2011a, b). Generally, all studies apply some approaches such as Fuzzy rule, AHP, adding some index variables and so on to develop a new and modified DRASTIC index. However, for the entire above-mentioned methods, in all studies vulnerability maps is implemented based on DRASTIC. Introducing a new methodology in providing vulnerability map based on intrinsic characteristic of each aquifer, leads to abounding of constant rank and weights. Here a clustering method is applied to achieve this goal.

Application of clustering, as one of the effective unsupervised datamining methods, in groundwater studies has been mostly reported for dealing with quantitative problems and groundwater quality issues specifically. For the first time, applications of clustering to groundwater research was introduced by Pedroli (1990) targeting classification of chemical composition of groundwater samples of Dutch shallow groundwater samples. Clustering techniques have been mostly focused on regionalization of a case study regarding the spatial and temporal variations of groundwater. In another similar study, clustering facilitated quality assessment of Dutch groundwater samples based on soil types and land use parameters (Frapporti et al., 1993). In a study, a combination of groundwater flow parameters and vegetation were selected as inputs of the clustering technique (Batelaan et al., 2003). Another interesting application of the clustering in groundwater studies, is determination ofgroundwater flow directions by applying the measured hydro-chemical concentrations of substances in water samples (Ochsenkuhn et al., 1997). The results showed that groundwater flow connections between the sample sites where similar water samples were taken from them have similar hydro-chemical characteristics (Riley et al., 1990). Along with successful application of the data mining method in groundwater studies, quality assessment of subsurface flow were studied. Different algorithms have been used to investigate new knowledge about existing quality dispersion patterns of the subsurface waters (Pedroli, 1990). Multi-objective fuzzy unsupervised pattern recognition approach is employed to assess potential pollution of subsurface flow area. Also a guidance for the industrial planning of groundwater sites was provided by Zhou, et al. (1999).

In this study, clustering technique is employed in regionalization of groundwater flow zone for the vulnerability assessment of a groundwater case study. To this, K-Means clustering as an unsupervised pattern recognition technique is applied. Thanks to the intelligent algorithm of clustering in finding similarities in the dataset, The proposed method of this research is capable of being used in each aquifer without considering calibration. The method is employed in a large-scale aquifer in the center of Iran, and the finding results are compared with vulnerability maps of the regions created by DRASTIC approach.

2. Methods and Materials

2.1. Study area

The Hashtgerd plain is located in the central part of Alborz Province covers an area of 410 km^2. The plain Extends from 50°29/- 51°6E to 35°47/- 36°07/ N (Fig. 1).

The area has a semi-humid climate in the northern part and a semiarid climate in the southern part. The main river of this plain is Kordan river which is becoming dry due to over use in the recent years. Fig. 1 shows the location of the case study and the quality of data samples.

2.2. Cluster Analysis

Clustering is an unsupervised pattern recognition technique, which tries to find

hidden structures out of a set of unlabeled data according to intrinsic similarities between the data (Hanand Kamber, 2006). Unlabeled data refers to datasets without any prior information in clustering process (Koskela, 2004). The similarity is computed mathematically in metric spaces and is defined by means of a distance norm (Van der Heijden et al., 2004).

2.3. Data Preparation

Dataset used for the proposed method in this study contains four features of depth of ground water; hydraulic conductivity; recharge value; and vadose zone for 4936 sample points. The topography of aquifer is relatively flat. Therefore, the topography factor is not considered in this study. The first three features have quantitative values, while the last one has a quality factor. Since the data set should contain quantity data, the vadose zone is transformed from quality form to quantity values of 1 to 5 for each data sample. Clay; Silt-Clay; Sand-Stone; Sand-gravel silt; and Sand-gravel types that constitutes different zones of the case study which is transformed to numbers 1 to 5 respectively. It should be mentioned that the applied features are the most important factors that are involved in almost entire groundwater vulnerability studies.

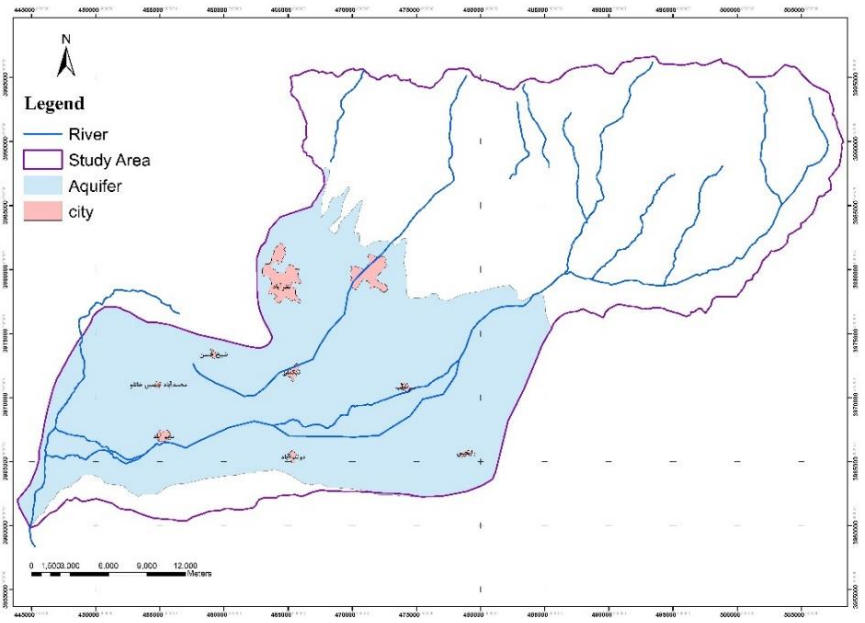

Fig 1. Location of Hashtgerd aquifer and quality data samples

2.4. K-means clustering

The process of grouping a set of abstract objects into classes of similar objects is called clustering. Cluster analysis, often used as a synonym for unsupervised pattern recognition and is applied to classify the set of unlabeled data. Unlabeled data refers to datasets that exist without any prior information for analyzing the data. Clustering techniques classify a data set according to resemblances between the data. The most important clue for the resemblance of two objects is the distance between the objects.

The K-means is a well-known partitional clustering methods, widely employed in different scientific fields, thanks to ease of implementation, simplicity and efficiency in application (Han and Kamber, 2006). K-means algorithm belongs to hard partitioning algorithms. In this algorithm a set of N data $(x_1, x_2, \ldots, x_N)$ in n dimensions are partitioned into c clusters, where each data point is allocated entirely to one cluster. It is an iterative process whereby the data are initially partitioned randomly, and iteratively reassigned to a cluster, based on the nearest distance to the clusters center. The procedure terminates when there is no reassignment of any data from one cluster to another (Weatherill and Burton, 2008). K-means algorithm partitions the $N \times n$ dimensional data matrix, X, into c clusters by minimizing the objective function (J) that is defined as equation 1 (Feil, 2006):

$$J(X;V)=\sum_{i=1}^{c}\sum_{k\in i}\left\|x_k^{(i)}-v_i\right\|^2 \quad (1)$$

Where X is the dataset, $V=\{v_i \mid i=1,...,c\}$ is the cluster centers and $x_k^{(i)}$ is the k^{th} object belonging to the i^{th} cluster. $\left\|x_k^{(i)}-v_i\right\|^2$ is distance measure norm indicating the distance between data points from their respective cluster centers that is calculated as follows:

$$v_i=\frac{\sum_{k=1}^{N_i}x_k}{N_i}, x_k\in A_i \quad (2)$$

Where A_i is the set of N_i number of objects belonging to i^{th} cluster. The process of clustering terminates when there is no reassignment of any data from one cluster to another (Weatherill and Burton, 2008).

One of the challenges in using a partition algorithm, like K-means, is choosing the optimal number of clusters (Weatherill and Burton, 2008). In this paper, a cluster separation index is used. This index is based on both the within-scatter of the clusters in a given clustering and the separation between the clusters. This measure is known as the DB Index (DBI) and proposed by Davies and Bouldin (Davies and Bouldin, 1979). The DBI value, in grouping n objects to g clusters is determined after computation of scores for all the possible pairs of clusters. The score is inversely proportional to the distance between the cluster centers and directly proportional to the sum of the within-scatters between every possible pairs of clusters. This score is given by equation 3 (Theodoridis and Koutroumbas, 2003):

$$R_{jk}=\frac{\sigma_j+\sigma_k}{\left\|\mu_j-\mu_k\right\|},\quad j,k=1,2,...,g;\ k\neq j \quad (3)$$

Here μ_j is the mean of all the objects in cluster j and σ_j is the within scatter of j^{th} cluster calculated as follows:

$$\sigma_j=\sqrt{\frac{1}{n_j}\sum_{x_i\in C_j}\left\|x_i-\mu_j\right\|^2} \quad (4)$$

Where the set of objects is C_j associated with cluster j and n_j is the number of objects in j^{th} cluster. The score is small when the means of j^{th} and k^{th} clusters are far apart. Since cluster j can be paired with $g-1$ other clusters, a conservative estimate of the cluster score for cluster j is obtained by assigning the maximal pair-score with cluster j:

$$R_j=\max_{k=1,2,...,g;k\neq j}R_{jk}. \quad (5)$$

The DBI of the complete clustering is then determined by averaging these maximal pair-scores for all clusters:

$$I_{DB}=\frac{1}{g}\sum_{i=1}^{g}R_j. \quad (6)$$

K-means clustering is carried out using the pattern recognition toolbox in Matlab prepared by Pattern Recognition Research Faculty of Delft University of Technology. Determination of the optimal number of clusters is carried on by running the K-means algorithm from two to maximum number of clusters (C_{max}). The value of C_{max} can be chosen according to the user's knowledge of the data set; however, as this is not always possible, a rule of thumb that many investigators use is $C_{max}\leq\sqrt{n}$, where n is the number of data (Kim et al., 2004). Two suggestions are presented about the optimal number of clusters in the literature. According to the first one, the optimal number of clusters occurs at the minimum DBI. The second suggests that, the optimal number of clusters occurs when value of the DBI becomes constant in the plot of DBI value versus number of clusters

3. Results and Discussion

3.1. Clustering Method

The clustering is sequentially operated from 2 to 15 numbers of clusters. The clustering separation index of DBI is computed in any iteration. Accordingly, the optimal number of clusters is equal to 4 number of clusters. The output of K-means clustering approach is summarized in tables 1 and also is depicted in Fig. 2

Table 1 gives the variation ranges of the applied features for the created clusters. By detailed observation on this table, it could be concluded that the entire features have been shared in data clustering. However, the features of depth and recharge show a dominant effect in clustering rather than the vadose zone and hydraulic conductivity features. Thus, a completely crisp boundary

between the clusters is created so that there is no similarity between each pairs of data from two different clusters. It stems from a more or less satisfactory clustering process that leads to complete separate clusters with no similarities from features, as depth and recharge. Also a more or less similar pattern happened for the hydraulic conductivity. The created clusters are ranked from lowest vulnerability to the highest one in table 1. Accordingly, cluster No. 1 is representative of the lowest vulnerable cluster that contains data with highest values of depth and lowest range of hydraulic conductivity and recharge. On the other hand, data with lowest values of depth and highest range of hydraulic conductivity and recharge constitute the most vulnerable cluster, i.e. cluster No. 4.

The vadose feature has similarities between the clusters. For instance clusters No. 1, No. 2, and No. 3 lay in vadose zone of "clay" and "silt clay". Clusters No. 4 with the highest risk of vulnerability is located in "sand stone", "sand gravel silt" and "sand gravel" zones. The range of features variation and especially depth to groundwater in the first and second one reveals that both of them could be considered as low vulnerability classes. Therefore, the sample points of these two clusters are merged together and lay in class of low vulnerability. Thereupon, according to Table 1 clusters No. 3 to No. 4 is respectively representative of "High", and "Very High" vulnerability. According to Fig. 2, the upstream regions where ground water table is near to the surface level and the hydraulic conductivity is more or less considerable, vulnerability degree is high and very high. However, the center and downstream parts of the case study shows low to medium degree of vulnerability.

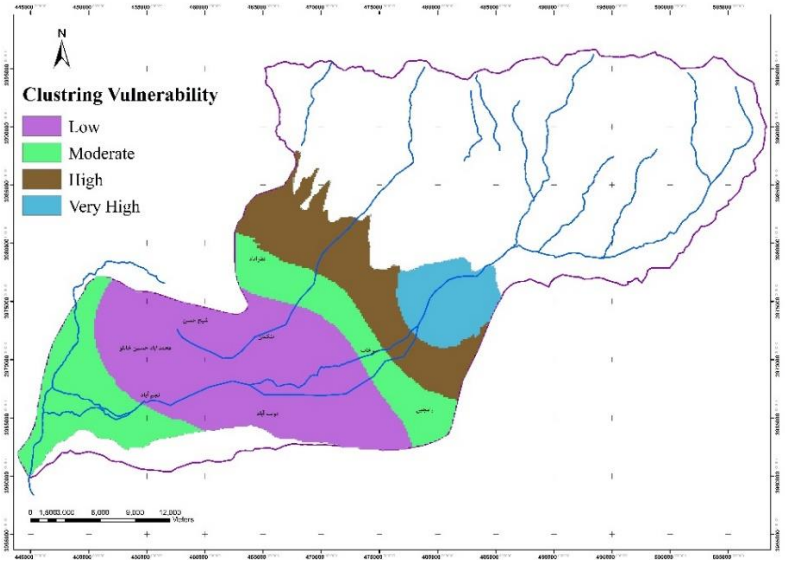

Fig. 2. Vulnerability maps by applying clustering methods

Table 1. Variation range of features in created clusters

Statistical Characteristic		Number of Data		Depth (m)		Hydraulic Conductivity (m/day)		Recharge (mm)		Vados Zone	
		Number	Percentage								
Raw Data		4253.00	100.00	min	max	min	max	min	max	min	max
Clusters	No. 1	1658.00	39	72.35	112.27	0.95	8.37	85.40	350.39	1	2
	No. 2	1148.00	27	48.60	72.29	6.38	15.45	320.45	800.42	1	3
	No. 3	936.00	22	35.12	48.56	15.80	20.50	800.42	1200.85	2	3
	No. 4	511.00	12	18.50	35.10	18.70	25.20	1200.37	1900.15	3	5

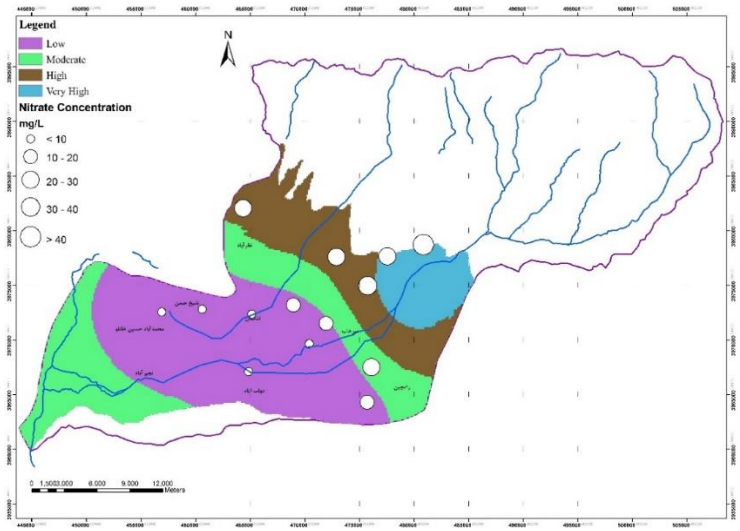

Fig. 3. Spatial distribution and amount of nitrate out of the samples points

3.2. Verification Methods

In this study, the verifications are carried out by measured samples of nitrate. Thereby, 52 number of nitrate samples, spread out all over Hashtgerd plain, are measured. The sampling time is selected after the period of fertilizer distribution. Spatial distribution of samples is demonstrated on the created vulnerability map by clustering method in Fig. 3. Pearson correlation coefficient is employed to figure out existing correlation between nitrate samples and vulnerability values obtained by clustering methods. The results show that application of clustering method leads to a reasonable correlation percentage (72%).

4. Conclusion

Vulnerability maps are applied as one of the effective management ways for qualitative management. Several models like DRASTIC were applied to this end and many researchers have tried to introduce approaches to provide more realistic constant ranks and weights using in the models. Meanwhile application of DRASTIC model is highly influenced by assigning the weights and ranks. Therefore, it is necessary to use a model that depends on variable weights and ranks according to the aquifer features.

In this paper clustering algorithm as one of the applicable data mining methods is used, taking the advantage of being independent from constant ranks and weights. In another word, vulnerability map of each region is provided by clustering according to the specific features of each region. Optimum number of clusters is obtained via 4 numbers by applying cluster validity index. Cluster map is then created based on spatial location of the created clusters representing the vulnerability map of the region. The map shows that central parts of the field are in high risk of groundwater vulnerability. Here, just four influential parameters are used by the clustering method in providing maps. Other parameters could be considered by clustering based on specific characteristics of each aquifer.

References

Aller, L., Bennet, T., Lehr, J. H., Petty, R. J. and Hackett, G., 1987, Drastic: a standardized system for evaluating groundwater pollution potential using hydrogeological settings, US Environmental Protection Agency.

Baalousha, H., 2006, Vulnerability assessment for the Gaza Strip, Palestine using Drastic, Environmental Geology, 50, 405-414.

Batelaan, O., De Smedt F. and Triest L., 2003, Regional ground-water discharge: phreatophyte mapping, groundwater modelling and impact analysis of land-use change, J. Hydrol., 275 (1-2) 86-108.

Davies, D. L. and Bouldin, D. W., 1979, A cluster separation measure, IEEE transaction on pattern analysis and machine intelligence, 1(4), 224-227.

Feil, B., 2006, Fuzzy clustering in process of data mining, PHD thesis, Department of Process Engineering, University of

Veszprem Hungry.

Frapporti, G., Vriend, P. and Van Gaans, P. F. M., 1993, Hydro-geochemistry of the shallow Dutch groundwater: interpretation of the national groundwater quality monitoring network, Water Resources Research, 29(9), 2993-3004.

Han, J. and Kamber, M., 2006, Data mining, concepts and techniques, San Francisco, U.S.A: Morgan Kaufman Publishers.

Javadi, S., Kavehkar, N., Mohammadi, K., Khodadi, A. and Kahawita, K. 2011a, Calibration Drastic using field measurements, sensitivity analysis and statistical method to assess groundwater vulnerability, Water International, 36, 719-732.

Javadi, S., Kavehkar, N., Mousavizadeh, M. H. and Mohammadi, K., 2011b, Modification of Drastic model to map groundwater vulnerability to pollution using nitrate measurements in agricultural areas, Journal of Agricultural Science Technology, 13, 239-249.

Kim, D. W., Lee, K. H. and Lee, D., 2004, On cluster validity index for estimation of the optimal number of fuzzy clusters. Journal of Pattern Recognition Society, 37(10), 2009-2025.

Koskela J., 2004, Pattern recognition in water resources management, A literature review and an application to long-term inflow forecasting, Master of Science thesis, Department of Civil and Environmental Engineering, Helsinki University of Technology.

Margat, J., 1968, Vulnerabilite des mappes d'eau souterraine a la pollution, Orleans: BRGM Publication.

Neshat, A., Pradhan, B. and Dadras, M., 2014, Groundwater vulnerability assessment using an improved Drastic method in GIS Resources, Conservation and Recycling, 86, 74-86.

Niknam, R., Mohammadi, K. and Majd, V. J., 2009, Aquifer vulnerability assessment using GIS and fuzzy system: a case study in Tehran-Karaj aquifer, Iran. Environmental Geology, 58, 437-446.

Nobre, R. C. M., Filho, O. C. R., Mansur, W. J., Nobre, M. M. M. and Cosenza, C. A. N., 2007, Groundwater vulnerability and risk mapping using GIS, modeling and a fuzzy logic tool, Journal of Contaminant Hydrology, 94, 277-292.

Ochsenkuhn, K. M., Kontoyannakos, J. and Ochsenku hn-Petropulu, M., 1997, A new approach to a hydrochemical study of groundwater flow, Journal of Hydrology, 194(1), 64-75.

Panagopoulos, G. P., Antonakos, A. K. and Lambrakis, N. J., 2006, Optimization of the Drastic method for groundwater vulnerability assessment via the use of simple statistical methods and GIS, Hydrogeology Journal, 14, 894-911.

Pedroli, B., 1990, Classification of shallow groundwater types in a dutch covers and landscape, Journal of Hydrology, 115, 361-375.

Riley, J. A., Steinhorst, R. K., Winter, G. V. and Williams, R. E., 1990, Statistical analysis of the hydrochemistry of ground waters in Columbia River basalts, Journal of Hydrology, 119(1-4), 245-262.

Saidi, S., Bouria, S., Dhiaa, H. B. and Anselmeb, B., 2011, Assessment of groundwater risk using intrinsic vulnerability and hazard mapping: application to Souassi aquifer, Tunisian Sahel, Agricultural Water Management, 98, 1671-1682.

Theodoridis, S. and Koutroumbas, k., 2003, Pattern recognition, Second edition. USA: Elsevier press.

Van der Heijden, F. and Duin, R. P. W. and De Ridder, D. and Tax, D. M. J., 2004, Classification, parameter estimation and state estimation, West Sussex, England, John wiley & sons Ltd.

Wang, J., He, J. and Chen, H., 2012, Assessment of groundwater contamination risk using hazard quantification, a modified Drastic model and groundwater value, Beijing Plain, China. Sci. Total Environ, 432, 216-226.

Weatherill, G. and Burton, P. W., 2008, Delineation of shallow seismic source zones using K-means cluster analysis, with application to the Aegean region, Geophysical Journal International, 176(2), 565-588.

Zhou, H., Wang, G. and Yang, Q., 1999, A multi-objective fuzzy pattern recognition model for assessing groundwater vulnerability based on the Drastic system, Hydrological Sciences Journal, 44(4), 611-618.