Earth Sci Inform (2015) 8:583–593
DOI 10.1007/s12145-014-0184-0

RESEARCH ARTICLE

Geoenvironmental zoning evaluation and optimization

Ahmad Zamani · Sedigheh Farahi · Reza Boostani · Hamed Hassani-Saadi

Received: 26 March 2014 / Accepted: 6 October 2014 / Published online: 19 October 2014

Abstract The problem of fast pattern classification by automatic analytical and sorting techniques is relevant across a wide range of scientific and technical disciplines. Since the availability of mass transactional and experimental data sets to address the challenges faced by the Earth and environmental scientists, validation of zoning has become an important topic. In this paper, we propose a new approach for producing Automatic Integrated Self-Organized Optimum Zoning (AISOOZ) maps using comprehensive (multivariate) geological and geophysical data. Unlike conventional zoning, the new approach includes techniques for finding the optimal structure that best fits natural pattern of a given area without the benefit of any a priori class information. While there are obvious similarities between the conventional and new optimal zoning maps, the automatic optimal approach reveals new insights into the geological evolution of the study area that could not be observed on the conventional maps. The success of the AISOOZ case study encourages the enlargement of its scope and application for rapid online as well as offline interactive multivariate pattern discovery in the Earth and environmental sciences studies. Finally, a comparative study between two widely used stopping criteria for optimal zoning and pattern recognition has been performed.

Keywords Tectonics · Neotectonics · Seismotectonics · Earthquake hazard · Classification · Silhouette index

Communicated by: H. A. Babaie

A. Zamani (✉) · S. Farahi
Department of Earth Sciences, College of Sciences, Shiraz University, Shiraz, Iran
e-mail: zamani_a_geol@yahoo.com

R. Boostani · H. Hassani-Saadi
Department of Computer Science and IT, Faculty of Engineering, Shiraz University, Shiraz, Iran

Introduction

One of the important capabilities of the human brain is its ability to classify entities based on the characteristics that the brain either recognizes or judges to be of importance. However, conventional methods of classification are characterized by two deficiencies. The first one is the large uncertainty involved in classification based on subjective analysis. Failure to visually interpret accurately a massive amount of data is the second. Automatic Integrated Self-Organized Optimum Zoning (AISOOZ) provides researchers with an empirical and objective method for performing this natural task for humans (Zamani and Hashemi 2004; Zamani et al. 2011, hereafter referred to as I, and II). Typically, the characteristics of the Earth and its natural phenomena are not only correlated with each other, but each characteristic is also influenced by other characteristics. Thus, in many instances the characteristics are interwoven in such a way that when analyzed individually, they produce little information about the Earth and its environment. Furthermore, the development of more sophisticated technology to collect information about the area under investigation has outpaced human's ability to use it to full potential. Today, Earth scientists have access to large quantities of data which contain far more useful information about the Earth and its environment than can be extracted by conventional methods. As the pace of decision making increases, researchers must also become more acquainted with new "field" of data analysis. Sifting through massive amount of data requires both statistical reduction and the ability to compute theoretical solution in Earth models with many parameters (I, II). Data mining (sometimes called data surfing or knowledge—discovery) is defined as the process of automatically collecting and searching through large amount of data in a database (Witten et al. 2011; Zamani et al. 2009, 2012, 2013b). As the term suggests, this new perspective of data analysis has a somewhat more exploratory rather than confirmatory nature. This technique is directed toward searching deeply into the characteristics of the large databases for patterns and relationships. Data mining uses artificial intelligence techniques, neural networks, and advanced statistical methods, to reveal trends, patterns, and relationships, which might otherwise have remained obscured by the complex patterns of association and massive amount of data. To do this, data mining uses computational techniques from statistics and pattern recognition (Zamani et al. 2012). Such data processing makes large volume of data digestible, discovers hidden rules underlying the data, and facilitates deduction of patterns and trends. Because of such properties, from a practical perspective, automated pattern recognition plays an essential role in many fields such as scientific data exploration, information retrieval and text mining, business, census, and education (Ashok Kumar and Kannathasan 2011). Among various data mining methods, statistical method of hierarchical classification is an important tool for "unsupervised" learning procedures—the problem of finding groups in data set without respect to external information. Unsupervised evaluation measures used in these procedures are often called internal indices because they use only information present in the data set. Whereas supervised learning procedures use some external evaluation measures often called external indices. In the context of data mining and pattern recognition hierarchical classification divides data set into groups by automatic analytical sorting techniques. Therefore, entities of same group are more similar than entities of different groups at different levels of partitioning. Nonhierarchical techniques assign objects or entities into groups once the number of groups to be formed is specified before any classification is accomplished. The procedure begins by selecting class seeds, which are used as initial guesses of the means of the classes. Class seeds can be specified by the researcher or selected randomly from all observations (i. e. objects or entities). Assuming *n* sampled sites in the study area, the first level is a partition of *n* sites into *n* groups each one containing one site. Therefore, at level *n* all sites form one group. If zones in the same groups stay together at higher levels, then the sequence is called a hierarchical classification (Zhu et al. 2008, 2010). As a result a series of nested partitions are generated, from the trivial partition with a single zone (containing all sampled sites) to the trivial partition where each sampling site is a zone. In between these two extremes are many candidate partitions or zonings that the researcher will have to choose from. The problem of determining the appropriate number of zones in an area is relevant across a wide variety of disciplines such as business, social sciences, biological sciences, Earth sciences, medicine, and engineering to name among others. Therefore, the use of a hierarchical classification method requires a decision regarding the stage or level which best reproduces the underlying structure. Since there is no internal statistical criterion used for inference, researchers have developed a number of criteria and guidelines for approaching the problem. These techniques that are sometimes referred to as stopping rules are ad hoc procedures and must be computed by the researchers. The resulting pattern or underlying structure could be considered to consist of "natural" or optimal zones in a given area. This underlying structure exhibits the properties of internal cohesion (compactness) of the sampled sites in a zone and external isolation (separation) of a zone from other zones (Halkidi et al. 2001; Tan et al. 2006). The interpretation of the resulting pattern depends very much on the correct choice of the number of zones. Therefore, to achieve such goal as selecting the "best" or optimal number of zones based on the distribution of zoning criterion, a stopping rule is desired (Chaimontree et al. 2010). Virtually all zoning procedures provide little if any information as to the number of zones present in the study area. Silhouette criterion is a zone validity index that is used to judge the quality of any zoning solution (partition). It reflects the relative similarity cohesion within zones and separation between zones. This criterion has gained popularity by showing more stability and robustness with the multivariate zoning classification. To illustrate the successful application of AISOOZ, Ward's classification method and Silhouette validation criterion were used to produce the optimal comprehensive (multivariate) geologic zoning (AISOOZ) map of Iran as a case study. While there are similarities between AISOOZ map and conventional (current) geologic maps, the comprehensive optimal zoning map reveals some remarkable features not found in the conventional maps. Finally, a comparative performance study of Wilk's lambda test statistic (II) and the Silhouette validation index for determining the optimum number of geological zones was conducted.

Method of analysis

Unsupervised hierarchical classification procedures are among the best known pattern recognition techniques (Grira et al. 2005). These procedures are needed to seek and separate zones with similar pattern in the dataset at different levels. In supervised classification, the validation of the resulting classification model is an integral part of the procedure of developing a classification model. However, because of its very nature, unsupervised zoning validation is not a commonly used part of classification analysis. Let us consider a hierarchical classification of *n* sampled sites or entities into *k* zones. The first level is a partition into *n* singletons, each one containing one site. The next level is a partition into *n*-one zones, and so on until all the sampled sites form one zone. At level *m* in the sequence the numbers of zone(s) is $c = n\text{-}m\text{+}1$. Therefore, level one corresponds to n singletons, each containing one site and level n in which all the sites form one zone (Zhu et al. 2008, 2010). Hierarchical classification is divided into agglomerative (bottom up or clumping) and divisive (top down or splitting) methods. Agglomerative methods begin with *n* singleton and the process is repeated by successively merging zones until all sites finally join together in a single zone. Divisive methods begin with one zone containing all the sampled sites, which in succeeding steps, split into smaller zones until each site is in a separate zone (singleton). These procedures are the opposite of the agglomerative procedures. Given a dataset, one of the major challenges to be solved in a classification process is to choose the classification model that could be more appropriate and explain better the underlying structure of data set (Ertöz et al. 2003). Another important objective in classification is to assess the "natural" or optimum level of partitioning in a given dataset, which is even more challenging when no response variable or classification model is available. Virtually all classification methods provide little if any information as to the optimum level of partitioning in the dataset. So, in this paper a new approach based on the application of Ward's minimum variance method for hierarchical classification and Silhouette validation criterion (also known as Silhouette width) is proposed to construct optimum geoenvironmental zoning maps. Ward's minimum variance method (Ward 1963) minimizes the loss of information with each geoenvironmental zone. Information loss is expressed in terms of a within-error sum-of-squares criterion, ESS, which is a built-in dissimilarity criterion (I). The Silhouette statistics is an internal criterion for unsupervised zoning validity, which combines both cohesion (compactness) and separation for determining classification evaluation (Fig. 1). It implies the assessment of zone coherence by comparing inter-to intra-zone variability (Halkidi et al. 2001; Tan et al. 2006). The cohesion and separation functions can be expressed using Eqs. (1) and (2), respectively (Jegatha Deborah et al. 2010; Tan et al. 2006).

$$cohesion\ (C_i) = \sum_{\substack{x\,\varepsilon\,c_i \\ y\,\varepsilon\,c_i}} proximity(x,y) \tag{1}$$

$$separation\ \left(C_i\ ,\ C_j\right) = \sum_{\substack{x\,\varepsilon\,c_i \\ y\,\varepsilon\,c_i}} proximity(x,y) \tag{2}$$

The evaluation function can be a similarity, dissimilarity, or some combination of these quantities. Equations 1 and 2 give us some well-defined measures of classification evaluation that can be combined into an overall measure of classification evaluation for a set of k zones, using a weighted sum of the validity of individual zones (Eq. 3).

$$overall\ validity = \sum_{i=1}^{k} w_i\ validity\ (C_i) \tag{3}$$

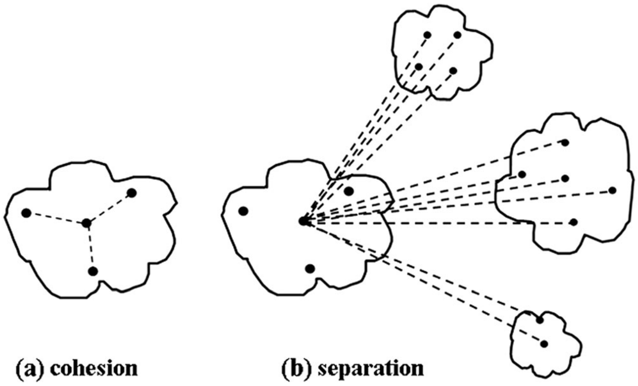

Fig. 1 Graphical representation of the cohesion and separation based on the Silhouette criterion (Jain and Dubes 1988; Tan et al. 2005)

The cohesion and separation validity functions used in the overall validation of a group of zones (Eq. 3) also can be applied to evaluate individual zones or objects. This information can be used to improve the quality of a classification process. The contribution of individual entities or sites within a zone also can be evaluated, using cohesion and separation validity functions (Jegatha Deborah et al. 2010; Tan et al. 2006). Objects that contribute more to the cohesion and separation are near the "interior" of the zone. On the other hand, items that contribute less to the cohesion are near the "edge" of the zone. The Silhouette validation criterion is a dimensionless stopping rule that has been used in a wide variety of disciplines for assessing the "natural" numbers of partitions (Rousseeuw 1987; Kaufman and Rousseeuw 1990, 2005). Its application for geoenvironmental zoning and patterns discovery in the Earth and environmental sciences is a novel approach to this problem.

Data analysis

To illustrate the successful application of AISOOZ, geological zoning map of Iran has been produced as a case study. For this purpose, large numbers of new and updated geophysical and geological attributes of Iran (Table 1) have been compiled for the 175 quadrangles sites of 1° areas.

The quadrangles from west to east are numbered beginning with 1 for the quadrangle between 44° E and 45° E meridians increasing to175 for the quadrangle between61° E and 62° E meridians (I, II). None of offshore Iran is included in the dataset. These quadrangles are used as items or observations (input samples). Each sampling site has been characterized by 49 attributes (Table 1) that seem to characterize the intensity and degree of contrast between tectonic, neotectonic, and seismotectonic characteristics of Iran. The attributes have all been standardized to have a mean of zero and a standard deviation of one. This removes the bias introduced by the

Table 1 Characteristics used for producing automatic integrated self-organized optimum zoning (AISOOZ) map, measured within 1° quadrangular sites[a]

No.	Geophysical attributes	No.	Geophysical and geological attributes
1	Minimum of gravity (mgal), MIGR	26	Maximum of magnetic intensity (gamma), MXMI
2	Maximum of gravity (mgal), MXGR	27	Average of magnetic intensity (gamma), AVMI
3	Average of gravity (mgal), AVGR	28	Range of magnetic intensity (gamma), RNMI
4	Range of Gravity (mgal), RNGR	29	Average of crustal thickness (km), ACTI
5	Minimum of free air anomaly (mgal), MIFR	30	a- value in the Gutenberg-Richter's formula, AVAL
6	Maximum of free air anomaly mgal), MXFR	31	b- value in the Gutenberg-Richter's formula, BVAL
7	Average of free air anomaly (mgal), AVFR	32	Number of earthquakes with magnitude 4.5 and greater, NEG4.5
8	Range of free air anomaly (mgal), RNFR	33	Number of earthquakes with magnitude smaller than 4.5, NES4.5
9	Minimum of bouguer anomaly (mgal), MIBO	34	Maximum earthquake magnitude, MXEM
10	Maximum of bouguer anomaly (mgal), MXBO	35	Earthquakes energy, EAEN
11	Average of bouguer anomaly (mgal), AVBO	36	Relative area of igneous rocks (%), RAIR
12	Range of bouguer anomaly (mgal), RNBO	37	Relative area of metamorphic rocks (%), RAMR
13	Minimum of regional bouguer anomaly (mgal), MIRG	38	Relative area of sedimentary rocks (%), RASR
14	Maximum of regional bouguer anomaly (mgal), MXRG	39	Relative area of ophiolit rocks (%), RAOR
15	Average of regional bouguer anomaly (mgal), AVRG	40	Relative area of unconsolidated sediments (%), RAUS
16	Range of regional Bouguer anomaly mgal),RNRG	41	Relative area of Cenozoic rocks (%), RACE
17	Minimum of residual bouguer anomaly (mgal), MIRS	42	Relative area of Mesozoic rocks (%), RAME
18	Maximum of residual bouguer anomaly (mgal), MXRS	43	Relative area of Paleozoic rocks (%), RAPA
19	Average of residual bouguer anomaly (mgal), AVRS	44	Relative area of Precambrian rocks (%), RAPR
20	Range of residual bouguer anomaly (mgal), RNRS	45	Fault length density (m^{-1}), FLDE
21	Minimum of isostatic anomaly (mgal), MIIS	46	Minimum of elevation (m), MIEL
22	Maximum of isostatic anomaly (mgal), MXIS	47	Maximum of elevation (m), MXEL
23	Average of isostatic anomaly (mgal), AVIS	48	Average of elevation (m), AVEL
24	Range of isostatic anomaly (mgal), RNIS	49	Range of elevation (m), RNEL
25	Minimum of magnetic intensity (gamma), MIMI		

[a] Geophysical and Geological data have been obtained from digitized and regular geological maps of Iran (geological Survey of Iran 2004). Seismological data were taken from earthquakes that occurred between the years 1900 up to 2011 (Engdahl et al. 2006; Gutenberg and Richter 1954; ISC 2011; NEIC 2011). Geophysical data have been taken from Dehghani and Makris (1984), World Digital Magnetic Anomaly Map (2007) and SRTM map (2008) and digital data from geological survey of Iran

differences in the scales of the attributes and ensures all have equal weight in the calculation of similarity. The normalized data set is then classified using Ward's minimum variance for hierarchical classification. This procedure which is more intuitive and computationally efficient minimizes information loss associated with zoning classification and creates small and even sized zones (I, II). At each step it merges those two elements, whose merging least increases their sums of squared differences from the mean. Finally, the unsupervised statistical measure, namely Silhouette index, which is based on the comparison of the compactness and separations, is here applied for the first time as an internal validation index for determining the optimum number of zones. The quantitative statistical Silhouette coefficient combines both cohesion and separation validity functions to evaluate individual sampling sites, zones, and the entire set of zones within the study region (Rousseeuw 1987; Kaufman and Rousseeuw 1990, 2005). The following steps explain how to perform zoning validation.

1 For each i_{th} site, let a_i be the average dissimilarity between i and all other sites within the same zone. We can interpret a_i as how well matched i is to the zone it assigned (the smaller the value, the better the matching).
2 For the i_{th} site and any zone not containing the site, calculate average dissimilarity of i to all the sites in the given zone. Repeat this for every zone of which i is not a member. Denote the lowest average dissimilarity to i of any such zone by b_i. This value can be seen as the dissimilarity between i and its nearest "neighbor" zone.
3 Finally, the Silhouette index of the i_{th} site (observation) is defined as:

$$SC_i = (b_i - a_i)/\max(a_i\ ,\ b_i) \tag{4}$$

This can be written as:

$$SC_i = \begin{cases} 1-a_i/b_i, & if \quad a_i < b_i \\ 0 \quad , & if \quad a_i = b_i \\ a_i/b_i-1 \,, & if \quad a_i > b_i \end{cases} \tag{5}$$

From the above definition it is clear that the Silhouette coefficient SC_i can vary between −1.0 and +1.0. If the Silhouette coefficient is close to 1.0, it means that sampling site is "well- partitioned". If the coefficient is close to zero, it means that site could be assigned to another closest zone as well (i. e. it is on a boundary between zones). If the coefficient is about −1.0, the sampled site is in the wrong zone.

Result and discussion

For a given number of zones, the global or overall average Silhouette coefficient for classification $\overline{SC}$ is simply the average of Silhouette index over all sampling sites i,

$$\overline{SC} = 1/n\sum_{i=1}^{n} SC_i \tag{6}$$

Where *n* is the number of observations (sampling sites).

The average Silhouette width is a measure of how tightly clustered all the sites in the zoning are. Thus the average Silhouette width of the entire data set is a measure of how appropriately the data has been partitioned and determines the optimum number of zones in a data set (Rousseeuw 1987; Kaufman and Rousseeuw 2005; Trauwaert et al. 1989). The largest overall average Silhouette score indicates a good zoning (Fig. 2). For current research, the 13-zone model with largest overall average Silhouette score is taken as the optimum number of zones. That is, the highest zone separability is achieved when we use 13 of them. The average Silhouette coefficient of a zone SC_j can be calculated by taking the average of Silhouette coefficients of sampling sites belonging to the given zone.

$$SC_j = 1/m_j\sum_{i=1}^{m_j} SC_i{}^j \tag{7}$$

Where m_j is the number of sampling sites within the j-th zone, *j*=1, 2, …, K zones.

Finally, the overall or global Silhouette coefficient for the full or entire zoning SC can be determined by calculating the average Silhouette coefficient of all zones (Petrović 2006).

$$SC = 1/k\sum_{j=1}^{K} SC_j \tag{8}$$

It is clear to see that both a zone's Silhouette and the global Silhouette take values between −1.0 and 1.0, both inclusive. The overall Silhouette index value compares all the sampling sites globally (Bizhani and Tarokh 2011). Therefore, local trends and outliers do not hinder locating the true overall average Silhouette coefficient. Pollard and van der Laan (2002) showed that the overall Silhouette index tends to be a global criteria in the sense that it is not necessarily maximized at the level of the partitioning which we would select visually but rather usually higher up in the model. In this study, the zone validity analysis is performed using the statistics toolbox designed for the cluster analysis written under Matlab (Kaufman and Rousseeuw 1990, 2005). As we mentioned, Silhouette method is a comprehensive method based on which we can both find the number of clusters and determine the association of each sample to the created clusters. It should be mentioned that Silhouette method enables us to correct a clustering process by evaluating such samples. More importantly, Silhouette method is not a clustering method; it is a robust evaluation method that gives an opinion about how well the samples are distributed in the clusters. A negative Silhouette value indicates that a sample is wrongly assigned to a cluster. Contrary, a maximized Silhouette value

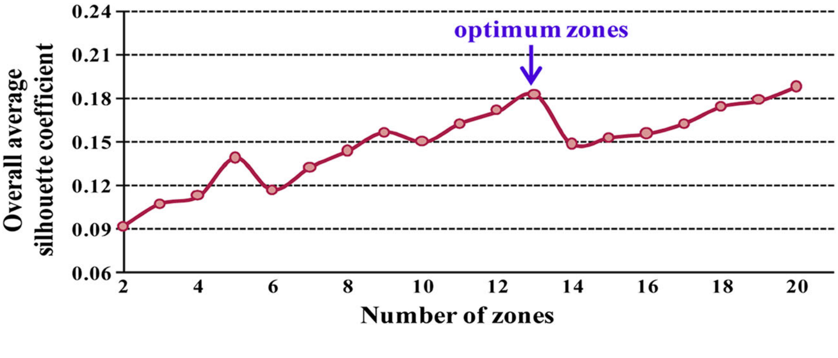

Fig. 2 Overall average Silhouette coefficient versus number of zones for the current research. There is a distinct peak in the Silhouette coefficient when the number of zones is equal to 13. This suggests 13-zone map is best, since it has the largest overall average Silhouette coefficient

demonstrates that a sample is properly assigned to a specific cluster. Our results show that, unlike other validation methods, the Silhouette method is not dependent on the employed clustering method. Therefore, the Silhouette method evaluates each sample individually while other methods evaluate the entire samples within a cluster as a group. You can produce a simpler and general classification defining the most basic structure of the area under investigation by comparing the zoning classifications globally. The Silhouette plot of the zoning (Fig. 3) shows the Silhouette index values across zones used to establish the 13 optimal geologic zones, each with specific numerical characteristics. It represents the Silhouette of all zones next to each other, so that the quality of zones can be compared.

The Silhouette index values for each zone is a plot of the Silhouette coefficients ranked in decreasing order of all the sampling sites within the same zone. The plot is a horizontal line, whose length is proportional to the Silhouette index value. The Silhouette plot shows which sampling sites lie well within the zone and which ones are somewhere in between zones. A wide Silhouette plot indicates large Silhouette index values and hence a pronounced zone. The height of a zone is simply equal to the number of sampling sites in the zone. The global or overall Silhouette width of 0.26 for the 13 optimal zones was found. As a rule of thumb, the Silhouette coefficient should be around or larger than 0.25 in order to be able to claim that a pattern in the data set has been found (Boroš 2011). Therefore, the Silhouette averages and plots can be used to select the optimal number of zones, assess how well individual observations are classified, and determine the underlying structure of the data set. As a case study of the successful application of the new approach, an optimal 13-zone geologic map of Iran is produced (Fig. 4). The geologic zones are numbered according to their hierarchical order. The following sections describe the features of each of these 13 optimal zones. The specific numerical geophysical and geological characteristics of these zones are presented in Table 2.

1) The Urumiyeh Zone: This zone is characterized by the moderate magnetic field intensity, moderate seismicity, relatively low earthquake energy release, moderate topographic relief, and moderate fault length density.
2) The Eastern Alborz - Kopeh Dagh Zone: Moderate exposure of Precambrian and metamorphic rocks, moderate magnetic field intensity and topographic relief, moderate seismicity, minimum area of Cenozoic rocks and high average elevation.
3) The Western Alborz Zone: This zone is distinguished from other zones by low crustal thickness, maximum outcrop of igneous rocks, and high gravity and isostatic anomalies.
4) The Hashtrud-Natanz Zone: Moderate magnetic field intensity and topographic relief, high percentage area of Precambrian, metamorphic, and igneous rocks exposures. Highest average elevation plays an important role in separating this zone from the other zones.
5) The Central Alborz Zone: The main geological characteristics of this zone are moderate fault length density,

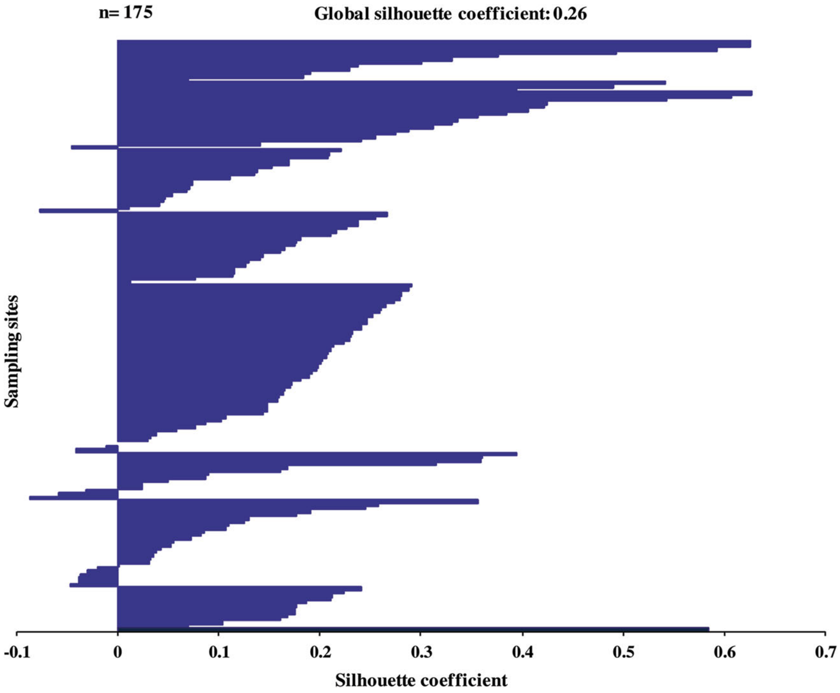

Fig. 3 The plot of Silhouette index values across zone used to establish the 13 optimal geologic zoning map of Iran. The Silhouette plot shows the sorted Silhouette indices (x-axis) for all sampling sites in each zone (y-axis). The global Silhouette coefficient of the zoning is the average of the zones' Silhouette widths

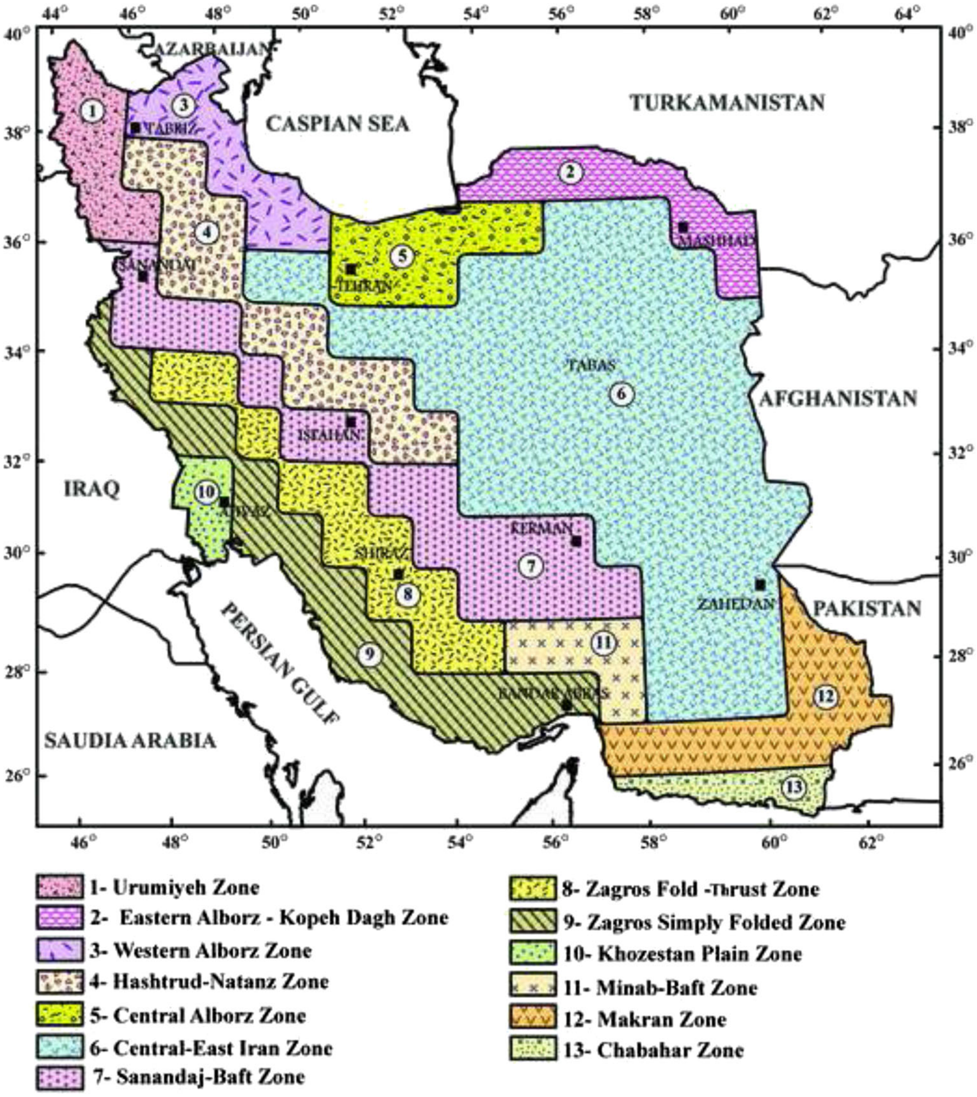

Fig. 4 Distinct Automatic Integrated Self-Organized Optimum Zoning (AISOOZ) of Iran based on Ward's minimum variance method for hierarchical classification and Silhouette validation criterion. Zoning has been made purely by reference to specified geophysical and geological characteristics shown in Table 1. The geologic zones are numbered according to their hierarchical order. The Corners of zones on map have been rounded slightly for cosmetic reason

maximum area of Paleozoic rocks exposures, high gravity anomaly, relatively high seismic severity, relatively high average Bouguer anomaly and low average magnetic field intensity.

6) The Central - East Iran Zone: This zone is differentiated from surrounding areas by the maximum area of Cenozoic rocks exposures, relatively low seismicity, low magnetic field intensity, moderate seismic severity and moderate unconsolidated sediments (Zamani et al. 2013a).
7) The Sanandaj - Baft Zone: Relatively high crustal thickness, high fault length density, moderate elevation, and moderate seismicity are important features of this zone.
8) The Zagros Fold - Thrust Zone: This zone is characterized by relatively low maximum isostatic anomaly (Zamani and Farahi 2011), high crustal thickness, moderate seismic activity and moderate elevation.
9) The Zagros Simply Folded Zone: This zone is differentiated from surrounding regions by moderate gravity anomaly, minimum outcrops of igneous and metamorphic rocks (Zamani and Farahi 2011), low magnetic field intensity and high seismicity. Maximum number of earthquakes occurred in this zone.
10) The Khozestan Plain Zone: This unfolded zone is characterized by low seismic activity, minimum fault length density, low isostatic anomaly, low topographic relief, and maximum exposures of unconsolidated sediments.
11) The Minab - Baft Zone: This zone is characterized by the moderate area of Paleozoic rocks, largest percentage area of ophiolite exposures, highest value of residual bouguer anomaly and magnetic field intensity.
12) The Makran Zone: Moderate regional bouguer anomaly, minimum area of Precambrian rocks exposures, maximum area of sedimentary rocks and high fault length density are the main characteristics of this zone.

Table 2 Geophysical and geological characteristics of 13 automatic integrated self-organized optimum zoning (AISOOZ) map shown in Fig. 4 (units and symbols are shown in Table 1)

Attributes	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6	Zone 7	Zone 8	Zone 9	Zone 10	Zone 11	Zone 12	Zone 13
MIGR	979306721	979427335	979579790	979152820	979249823	979165737	978856980	978837537	979002181	979340973	978722797	978753180	978992228
MXGR	979645252	979510806	980014490	979428729	979882363	979334798	979133253	979085220	979180479	979367663	978994050	978963559	979039690
AVGR	979492369	979433484	979772512	979301026	979585610	979260574	979004377	978958428	979118957	979353134	978901123	978848930	979019006
RNGR	338531.538	175032.429	434700	275909	632540	169060.816	276272.857	247682.632	178298.571	26690	271253.333	210378.75	47462.5
MIFR	−4106.9231	−4753.1429	−5631.6667	−1670	−9825	−3161.2245	−2126.6667	−5115.2632	−9313.5714	−9406.6667	−10460	−3337.5	1117.5
MXFR	13106.1538	5004.57143	13818.3333	11401	15975	6092.44898	14494.2857	6864.21053	206.428571	−6160	14400	5001.25	6177.5
AVFR	2794.63201	−95.788567	5501.75629	4074.36986	1969.32031	903.415618	4093.05673	539.058095	−5759.9265	−7595.4687	1216.98935	805.880013	3467.76291
RNFR	17213.0769	9757.71429	19450	13071	25800	9253.67347	16620.9524	11979.4737	9520	3246.66667	24860	8338.75	5060
MIBO	−17503.077	−11375.429	−12890	−18198	−20112.5	−13189.184	−18853.333	−20183.684	−11880.714	−9370	−15916.667	−11683.75	−110
MXBO	−8943.0769	−6391.7143	2401.66667	−11058	−650	−7002.449	−10123.333	−12567.895	−6816.4286	−6330	−3363.3333	−2220	5582.5
AVBO	−12898.675	−8514.5762	−5345.0345	−14693.835	−8595.4246	−10109.183	−14960.089	−16391.412	−9429.2711	−7702.7087	−9475.0172	−7946.0384	2992.64011
RNBO	8560	4983.71429	15291.6667	7140	19462.5	6186.73469	8730	7615.78947	5064.28571	3040	12553.3333	9463.75	5692.5
MIRG	−132.61538	−98	−84.166667	−158.4	−78.5	−117.30612	−156.19048	−165.52632	−149.71429	−111.66667	−141.33333	−126.125	−21.75
MXRG	−83.384615	−82.142857	−8.3333333	−110	−42.5	−96.081633	−128.61905	−136.47368	−79.357143	0	−103.33333	−32.625	82
AVRG	−108.83538	−89.797143	−46.533333	−134.938	−61.7425	−106.53143	−142.2181	−151.33579	−114.33143	−56.536667	−122.97	−79.45875	29.095
RNRG	49.2307692	15.8571429	75.8333333	48.4	36	21.2244898	27.5714286	29.0526316	70.3571429	111.666667	38	93.5	103.75
MIRS	−45	−14.428571	−43.166667	−34.9	−53.25	−17.387755	−19.380952	−33.157895	−11.857143	−66	−7.3333333	−8.125	−16.25
MXRS	23.9230769	20.0714286	33.6666667	28.8	38.75	30.7142857	25.5238095	20.5263158	40.6428571	7.66666667	89	63	40.5
AVRS	−9.2584615	2.76714286	−4.7493333	−2.911	−7.485	6.62367347	3.76952381	−6.91	16.0514286	−28.54	40.1666667	28.355	12.35
RNRS	68.9230769	34.5	76.8333333	63.7	92	48.1020408	44.9047619	53.6842105	52.5	73.6666667	96.3333333	71.125	56.75
MIIS	−4.9230769	−7.4285714	29.1666667	−1.4	7	−18.265306	−7.1428571	−27.315789	−64.357143	−79.333333	−29.333333	−28.75	24.25
MXIS	34.8461538	13	79.8333333	35.2	50	5.48979592	23.1904762	13.7368421	1.21428571	−58.666667	74	30.125	63.75
AVIS	15.76776	4.20255102	53.4488095	18.2763889	27.5059524	−4.4305879	10.8664399	−6.7664331	−32.356293	−65.496672	21.9825708	0.09652778	45.1577381
RNIS	39.7692308	20.4285714	50.6666667	36.6	43	23.755102	30.3333333	41.0526316	65.5714286	20.6666667	103.333333	58.875	39.5
MIMI	−165.07973	−111.22143	−159.78	−85.154	−173.145	−174.51265	−152.2981	−59.106316	−48.829286	−20.9	−121.69333	−23.64875	−82.025
MXMI	274.756923	135.347143	386.72	178.208	86.365	207.160408	235.994762	53.2247368	18.8635714	12.2766667	429.273333	161.54875	50.8675
AVMI	−12.628699	10.2403798	55.7590249	4.26340136	−48.51678	−1.3895909	1.20761391	−6.3331758	−12.836861	−5.3026606	81.1478685	33.8039427	−0.9011508
RNMI	457.28	246.568571	546.5	263.362	259.51	381.673061	388.292857	112.331053	67.6928571	33.1766667	550.966667	185.1975	132.8925
ACTI	42.7315385	39.8957143	32.8233333	45.636	38.67	41.4918367	44.287619	47.4847368	43.2278571	39.4933333	38.9666667	37.97	28.25
AVAL	4.19358974	4.227	5.11166667	4.29893333	4.235	4.0837932	3.96344603	5.05647368	5.45857143	6.252	5.52666667	5.637625	6.015
BVAL	1.06801026	1.04697143	1.26333333	1.07930667	1.0565	1.0436851	1.00562317	1.17222632	1.203	1.47733333	1.20666667	1.320225	1.42525
NEG4.5	9.84615385	14.2142857	16.3333333	9.8	18.5	7.3877551	5.47619048	25.3684211	73.2857143	2.66666667	61.3333333	7.625	2.25
NES4.5	11.3076923	14.5714286	13.5	12.4	16.25	8.79591837	7.47619048	42.6842105	110.428571	7.33333333	76.6666667	5.5	5
MXEM	5.43076923	5.46428571	5.53333333	5.58	6.075	5.27346939	5.07142857	5.47894737	5.74285714	4.7	5.8	5.625	5.75
EAEN	1.8807E+18	6.2811E+20	2.1477E+18	5.1908E+21	9.1437E+19	7.395E+19	7.8104E+19	5.429E+18	1.7682E+18	3.4441E+16	2.0929E+18	2.6567E+20	6.4018E+20
RAIR	9.85852029	0.47178472	29.1310619	6.87296	6.29970948	11.3608368	9.44837225	1.33282517	0	0	7.08857419	0.41190853	1.78413126

Table 2 (continued)

Attributes	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6	Zone 7	Zone 8	Zone 9	Zone 10	Zone 11	Zone 12	Zone 13
RAMR	1.97819313	0.11183986	1.22044643	16.8715769	1.77989551	1.90155203	2.2000503	1.69899078	0	0	15.8950165	2.13075687	6.32697287
RASR	46.1791366	32.5868776	21.8441208	39.577964	54.5808651	25.4419854	34.7661841	58.9268332	55.3117275	2.59183515	25.342807	66.4410221	33.9729312
RAOR	0.45266245	0.04419069	0.02101301	2.54805263	0	1.81260103	0.73822408	2.20800471	0	0	10.8876035	4.57511713	0
RAUS	36.6279592	17.9823428	19.5779117	23.1133843	37.3395299	57.7244472	52.1681706	35.8333461	40.9251088	67.4968676	40.7859988	17.263294	49.3967905
RACE	71.8214233	26.6905476	61.132241	41.4124516	43.764189	87.0839133	81.2407737	66.9565067	80.8757421	68.0836523	70.2641292	83.0634163	57.0991154
RAME	15.0748525	21.3589927	7.19679717	33.8701916	27.0392694	8.98624397	14.5463931	30.1468601	5.44408738	0	20.6765561	6.22781467	0
RAPA	3.43251845	0.18117611	0.60041822	3.38440404	11.8226704	0.9141108	2.27231455	1.38894425	0.32743724	0	6.04816098	0.49139223	0
RAPR	1.11169828	0.00013773	0.29886392	10.3079685	2.87776576	0.68967834	1.255427	0.79023894	0.32743724	0	3.0111537	0	0
FLDE	0.00572976	0.00340528	0.00346215	0.00654358	0.01461988	0.00540403	0.00625005	0.00618565	0.00251604	0.00029952	0.01076353	0.01844929	0.00438053
MIEL	703.398318	262.57675	87.6684772	996.303627	1.28266012	645.165343	1158.9121	909.956852	23.2110888	1.04152548	434.549624	254.306507	0.26501333
MXEL	3458.81105	2123.17347	2670.30486	3282.25066	3873.87295	2317.58444	3216.24094	3367.28093	2185.99504	145.250364	2929.52692	2025.45845	740.719863
AVEL	1551.34209	762.976195	1094.79382	1875.38187	1665.42245	1047.32743	1782.58625	1773.76051	543.72046	13.0654036	1148.47197	890.670997	107.672842
RNEL	2755.41273	1860.59672	2582.63638	2285.94704	3872.59029	1672.4191	2057.32884	2457.32408	2162.78395	144.208839	2494.9773	1771.15194	740.454849

13) The Chabahar Zone: Relatively low elevation and highest value of regional bouguer anomaly, low crustal thickness, low seismicity, and minimum percentage area of Mesozoic rocks exposures are the general characteristics of this zone.

Although there are some similarities between AISOOZ and conventional maps, AISOOZ map (Fig. 4) reveals some remarkable features that could not be found in the conventional maps (i.e. Stöcklin 1968; Stöcklin and Nabavi 1973). These conventional maps which are not only original but also widely used in Iran have been considered for illustrative purposes only. New features revealed by the robust multivariate geological zoning (AISOOZ) method (Fig. 4) include the following: 1- Contrary to the current maps AISOOZ map makes a clear distinction between the Makran Mountains (Zone 12) and the East Iranian Ranges. 2- The extent and rigidity of the disputed microplate in central and eastern parts of Iran (Zone 6) is revealed. 3- The robust multivariate zoning method displays striking similarities between the Eastern Alborz - Kopeh Dagh Zone (Zone 2) and the Zagros Simply Folded Zone (Zone 9) even though these zones are geographically far apart. They probably have gone through similar geologic processes during their geological evolution. 4- On the other hand, the Sanandaj - Baft Zone (Zone 7) and the Zagros Fold - Thrust Zone (Zone 8) that are geographically close are geologically differentiated from each other. This indicates that these zones went through separate paths during their geologic evolution.

What we are primarily concerned with in this manuscript is the mechanics of producing Automatic Integrated Self-Organized Optimum Zoning (AISOOZ) map based on overall similarities from a list of variables within individual zones. This new approach has been applied to zoning tectonic map for illustrative purpose only. It is important to recognize that the geological zoning generated by AISOOZ method is based purely on the geological and geophysical characteristics presented in Table 1. So, the correspondences and differences between AISOOZ and a given zoning based on conventional methods must receive careful thought (I & II).

However, the attributes used could readily be adapted to take additional or alternative data sets. The same approach could also be used to zoning maps of any entity, regardless of its context and/or scale.

Assessment of geoenvironmental zoning

Apart from primary validity, which assures that the zoning classification is representative of the general population, there is a measure of secondary validity. It assesses how well the classification is stable and robust. According to Williams (1967), a classification should be truly stable and robust in that slight alterations of information should not produce major changes in the classification. The clustering is naturally a “blind” process if it does not include any explicit semantic knowledge of the dataset. The most direct approach to evaluating and testing the stability and robustness of the Silhouette validation index is to alter the data set slightly by adding or removing of a few data samples and repeat the classification. For this purpose, the Silhouette validation criterion has also been applied to the data set used in our previous paper (II). This is slightly different than the up-to-date data set with the current research (Table 1). Since the procedure produced the same zoning classification on repeated trials (Fig. 4), one can conclude that the Silhouette validation index is a reliable measure for determining the optimum number of zones. Finally, the same reliability testing was performed on the Wilk’s lambda validation index (II). The resulting plots are presented in Fig. 5.

Comparison of Fig. 5(a) and (b) indicates that the Wilk’s lambda criterion tends to produce a different number of zones with a slight modification of the data set. This is a serious error because information is lost by merging distinct zones. Therefore, in comparison with the Wilk’s lambda criterion, Silhouette index performs better for optimum geoenvironmental zoning because it produces more stable and robust zoning classifications.

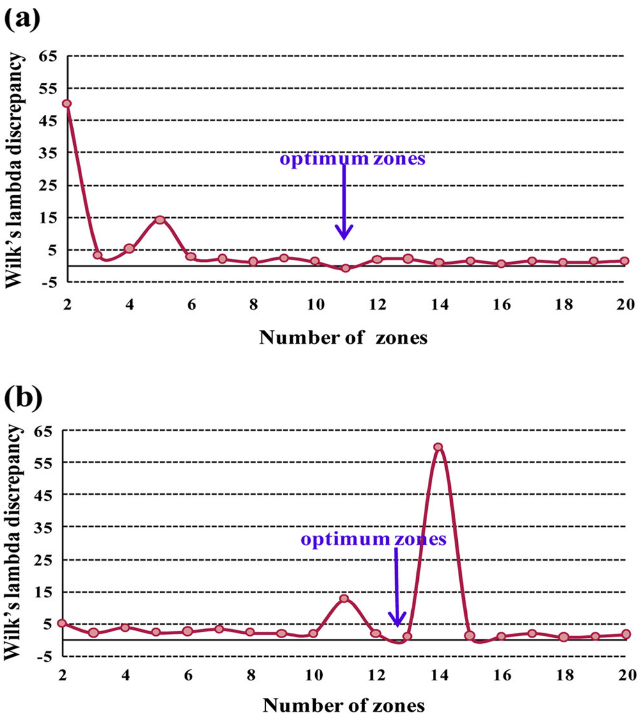

Fig. 5 Showing how the application of Wilk’s lambda criterion as stopping rule may result in different optimal number of geologic zones due to the slight alteration of the data samples **a** Wilk’s lambda discrepancy, calculated for the data set used in our previous paper (Zamani et al. 2011). **b** Wilk’s lambda discrepancy calculated for the up-to-date data from current research. This indicates that in comparison with Wilk’s lambda stopping criterion the Silhouette index is a more reliable and robust criterion

Conclusion

Zoning classification provides us with an empirical and objective technique for performing one of the most inherent capabilities for the human- classification. It aims at identifying groups of similar entities or sampling sites and helps to discover distribution of patterns and interesting correlations in the study area. Whether for purposes of simplification, exploration or confirmation, zoning classification is a powerful analytical method that has a wide range of applications in virtually every scientific and technical discipline. Since the availability of massive transactional and experimental data sets, validating the zoning results has become an important topic in the context of the Earth and environmental sciences. Having decided on a suitable classification method, consideration needs to be given to how to find a partition into groups that optimizes the number of zones. Because there is no internal statistical criterion used for inference, researchers have developed several criteria and guidelines for approaching the problem (Everitt and Dunn 2010). These criteria are ad hoc procedures and must be computed by the researcher after the experiment is completed. When applied to the results of zoning classifications, these procedures are sometimes referred to as stopping rules. In this paper Ward unsupervised classification method is used to produce geological zoning maps. After that, the Silhouette validation criterion has been applied for the first time as stopping rule to measure the relative usefulness of geological zoning maps. By comparison the sampling site globally and ignoring local dependencies, the average Silhouette index of the zoning classifications as a whole can determine the optimal or natural number of geologic zones within the study area. This criterion, eventually led to an AISOOZ map displaying 13 optimal geologic zone map of Iran. Finally, a comparative study has been performed on the geological zoning classifications produced by the Silhouette validation coefficient and that of the Wilk’s lambda criterion. In comparison with the Wilk’s lambda, the geological zoning produced by the Silhouette criterion is more reliable in that the resulting geological classification is not disturbed by the adding or removing a few sites or attributes to the data set. Therefore, this criterion could be built into computer programs as validation index or stopping rule to assess natural or optimum number of zones, each with specific numerical characteristics. AISOOZ maps not only can be reevaluated quite often, but also provides us with a means to create online as well as offline interactive zoning maps in the Earth and environmental sciences. The attributes utilized could easily be adapted to take additional or alternative data sets. Therefore, AISOOZ approach could also be used to zoning maps of any entity, regardless of its context and/or scale.

Acknowledgments This work was supported by the Center of Excellence for Environmental Geohazards, and the Research Council of Shiraz University. One of the authors (S.F.) is grateful to the Ministry of Science, Research and Technology of Iran and Iranian National Institute of Oceanography for financial support.

References

Ashok Kumar D, Kannathasan N (2011) A survey on data mining and pattern recognition techniques for soil data mining. Int J Comput Sci 8:422–428

Bizhani M, Tarokh MJ (2011) Behavioral rules of bank's point-of-sale for segments description and scoring prediction. Int J Ind Eng Comput 2:337–350

Boroš M (2011) Cluster analysis, methodology and statistics. University of Groningen. World Wide Web Address: WWW.let.rug.nl/~nerbonne

Chaimontree S, Atkinson K, Coenen F (2010) Best Zoneing configuration metrics: towards multi- agent based zoning. ADMA, proceedings of the 6th in international conference on advanced data mining and applications: part 1. Springer 6440:48–59

Dehghani GA, Makris J (1984) The gravity field and crustal structure of Iran. N Jahrbuchfür Geol Paläontol 168:215–229

Engdahl ER, Jackson JA, Myers SC, Bergman EA, Priestley K (2006) Relocation and assessment of seismicity in the Iran region. Geophys J Int 167:761–778

Ertöz L, Steinbach M, Kumar V (2003) Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. International Conference on Data Mining, San Francisco, California, USA

Everitt BS, Dunn G (2010) Applied multivariate data analysis, 2nd edn. Wiley, New York, p 354

Grira N, Crucianu M, Boujemaa N (2005) Techniques for processing multimedia content. Report of the MUSCLE European Network of Excellence (6th Framework Programme)

Gutenberg B, Richter CF (1954) Seismicity of the earth and associated phenomena. Princeton University Press, New Jersey

Halkidi M, Batistakis Y, Vazirgiannis M (2001) On clustering validation techniques. J Intell Inf Syst 17:107–145

ISC: International Seismological Centre (2011) On-line Bulletin. http://www.isc.ac.uk, Internatl. Seis. Cent., Thatcham, United Kingdom

Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall, Inc

Jegatha Deborah L, Baskaran R, Kannan A (2010) A survey on internal validity measure for cluster validation. Int J Comput Sci Eng Surv (IJCSES) 1(fir):85–102

Kaufman L, Rousseeuw PJ (1990) Clustering large data sets. In: Gelsema ES, Kanal LN (eds) Pattern recognition in practice II. Elsevier, North Holland, pp 425–437

Kaufman L, Rousseeuw PJ (2005) Finding groups in data: An introduction to cluster analysis. Wiley-Interscience, New York, p 368

National atlas of thematic geological & exploration maps (2004) Geological Survey of Iran, Ministry of Industries and Mines

NEIC: National Earthquake Information Center (2011) Earthquake information bulletin Colorado, USA. http://earthquake.usgs.gov/regional/neic/

Petrović S (2006) Proceedings of the 11th Nordic workshop on secure IT-systems. NORDSEC, Linkoping, pp 53–64

Pollard KS, van der Laan MJ (2002) A method to identify significant clusters in gene expression data. Paper 107, division of biostatistics working paper series. University of California, Berkeley

Rousseeuw PJ (1987) Silhouettes, a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65

SRTM: Hole-filled seamless SRTM data V, International Centre for Tropical Agriculture (CIAT) available from http://gisweb.ciat.cgiar.org/sig/90m_data_tropics.htm (2008)

Stöcklin J (1968) Structural history and tectonics of Iran: a review. Am Assoc Petrol Geol Bull 52:1229–1258

Stöcklin J, Nabavi MH (1973)1⁄2,500,000 sheet, tectonic map of Iran. Geological Survey of Iran

Tan PN, Steinbach M, Kumar V (2005) Introduction to Data Mining. Addison-Wesley; 1 edition, ISBN-13: 9780321321367

Tan P, Steinbach M, Kumar V (2006) Introduction to Data Mining. Addison-Wesley, pp. 769

Trauwaert E, Rousseeuw PJ, Kaufman L (1989) Some silhouette-based graphics for clustering interpretation. Belg J Oper Res Stat Comp Sci 29(3):35–55

Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244

Williams WT (1967) Numbers, taxonomy, and judgment. Bot Rev 33: 379–386

Witten IH, Frank E, Hall MA (2011) Data mining, practical machine learning tools and techniques. 3rd ed., Elsevier, pp. 629

World Digital Magnetic Anomaly Map (WDMAM) (2007) UNESCO Commission

Zamani A, Farahi Ghasre Aboonasr S (2011) The significance of parameters used for self-organized tectonic zoning of Iran. Sci Q J Geosci 79:165–170

Zamani A, Hashemi N (2004) Computer-based self-organized tectonic zoning: a tentative pattern recognition for Iran. Comput Geosci 30: 705–718

Zamani A, Nedaei M, Boostani R (2009) Tectonic zoning of Iran based on self organizing map. J Appl Sci 9:4099–4114

Zamani A, Khalili M, Gerami A (2011) Computer-based self-organized tectonic zoning revisited: Scientific criterion for determining the optimum number of zones. Tectonophysics 510(1):207–216

Zamani A, Sami A, Khalili M (2012) Multivariate rule—based seismicity map of Iran: a data—driven model. Bull Earthq Eng 10:1667–1683

Zamani A, Samiee J, Kirby JF (2013a) Estimating the mechanical anisotropy of the Iranian lithosphere using the wavelet coherence method. Tectonophysics 601:139–147

Zamani A, Sorbi MR, Safavi AA (2013b) Application of neural network and ANFIS model for earthquake occurrence in Iran. Earth Sci Inform 6(2):71–85

Zhu L, Lin C, Huang H, Chen Y, Yuille AL (2008) Unsupervised structure learning: hierarchical recursive composition, suspicious coincidence and competitive exclusion. Proc. Of the European Conference on Computer Vision (ECCV)

Zhu L, Chen Y, Yuille A (2010) Latent hierarchical structure learning for object detection. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)