Bull Earthquake Eng (2012) 10:1667–1683
DOI 10.1007/s10518-012-9386-3

ORIGINAL RESEARCH PAPER

Multivariate rule-based seismicity map of Iran: a data-driven model

Ahmad Zamani · Ashkan Sami · Marziyeh Khalili

Received: 25 June 2012 / Accepted: 27 September 2012 / Published online: 6 October 2012

Abstract The seismic hazard map or delineation of regions with high earthquake hazard is important to plan risk mitigation strategies. Identifying areas of high seismic hazard can lead city planners to enforce better construction standards and predict areas vulnerable to slope instability. Conventional seismic hazard maps are based on limited factors like ground acceleration, ground velocity, etc. This paper presents a new class of data-driven multivariate rule–based model to create online as well as offline interactive seismic hazard map that is flexible and readily automated. A multivariate rule-based seismicity map (MRBSM) is defined as the map of regions with a future high hazard of earthquakes. The classification and regression tree method is used to extract rules that predict regions with high hazard of earthquakes with $m_b \geq 4.5$ in Iran. The rules generated for our MRBSM of Iran are based on a large number of geological and geophysical parameters. The MRBSM indicates that the province of Bandar Abbas, a major population center in the South of Iran has a high hazard of earthquakes with $m_b \geq 4.5$. In addition, our method allows identification of the most important parameters associated with earthquakes. Our analysis shows that the isostatic anomaly has the strongest correlation with earthquakes while magnetic intensity, regional Bouger anomaly, Bouger anomaly, and gravity anomaly also correlate well. Despite widespread application of a- and b-values of the Gutenberg-Richter formula, these parameters do not correlate well with earthquake hazards in the area.

Keywords Data mining · Decision tree · Neotectonics · Seismotectonics · Earthquake hazard prediction · Iran

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

A. Sami
Department of IT and Computer Engineering, Shiraz University, 71347-51154 Shiraz, Iran

1 Introduction

Iran is one the most seismically active areas of the world and frequently suffers from destructive earthquakes that leave large numbers of casualties and financial losses. Earthquakes in Iran and neighboring countries are closely related to their positions within the geologically active Alpine-Himalayan Belt that separates the Eurasian from the Africa plates. Tectonic activity in this region can be characterized by high topography, recent volcanism and many active faults that cause destructive earthquakes. This area of complex plate interaction has long been the focus of attention among Earth scientists. Many researchers have studied the seismicity of Iran for decades. For example, Nowroozi (1976, 1979) introduced 23 seismotectonic provinces for Iran based on seismicity data, geological information, physiographic features, structural trends, active faults and distribution of salt domes. Berberian (1979) criticized this seismotectonic zoning. Shoja-Taheri and Niazi (1981) divided the country into three major seismic zones based on the seismic strain release by earthquakes. Ambraseys and Melville (1982) defined four major zones of seismic activity in Iran based on historical macroseismic data. The four proposed zones depict an overall pattern of seismic distribution in the country and were constructed without prior knowledge of regional tectonics. Karakaisis (1994) divided Iran into 21 seismogenic source areas based on the meizoseismal regions of destructive earthquakes, and major faults of Quaternary and Tertiary age. Tavakoli (1996) divided Iran into 20 seismotectonic provinces Tavakoli and Ghafory-Ashtiany (1999) assessed seismic risk in Iran. Bonini et al. (2003) studied the seismotectonic pattern of Iran using analogue models. Recently Ashtari Jafari (2010) used a statistical method to predict great earthquakes in Tehran.

This work deploys both surface and sub-surface data (geological and geophysical characteristics) to build a multivariate numerical database. Then rules governing high impact earthquakes were extracted based on combinations of major parameters. A decision tree rule extraction (data mining) method was used to shed light on the seismicity patterns and to determine the parameters that correlate highly with earthquakes in Iran. The application of data mining and machine learning methods such as neural network (Fu 1999) and decision tree (Quinlan 1993) is very common in a variety of fields that include the environmental sciences (Dmeroski 2002). While Rule-Based methods are not new, their application to earthquake hazard prediction is novel. A major advantage of using such methods is that they are mostly data driven, nonparametric and without priori assumptions. When models are based on data alone without any discrimination based on the researchers' opinions, historical facts are the main players in model construction. During the last few years, some researchers used machine-learning methods to build classifiers or to predict earthquakes. For example, Zmazek et al. (2003) used a regression tree (model tree) to predict earthquakes based on soil radon data. Iftikhar and Toshinori (2009) used a rough set and decision tree (C4.5 algorithm) to characterize premonitory factors of low seismic activity. Standart et al. (2010) applied data mining techniques to the discovery of spatial and temporal earthquake relationships. None of these studies has shown a combination of parameters that lead to identification of highly active seismic areas. Previous studies such as: (Berg et al. 1964; Bouchon 1973; Davis and West 1973; Caputo et al. 1984, 1985; Geli et al. 1988; Johnston 1997; Zamani and Hashemi 2000; Chen et al. 2002; Li and Li 2009) indicate that the seismicity of an area is influenced by geological and geophysical parameters such as isostatic anomaly, topography, gravity anomaly and the electromagnetic field. So, the rules governing the patterns of earthquakes and the relative importance of such governing factors have not yet been investigated.

This paper introduces for the first time the classification and regression tree (CART) analysis to extract meaningful rules that can be used to build a decision support system that predicts earthquake hazards.

CART, introduced by Breiman et al. (1984), has been applied fruitfully for both prediction and rule extraction problems. CART analysis is a machine learning method based on statistical rules that finds combinations of predictor variables that predict target variables. In other words, CART analysis is used to predict numerical outcome based on different variables. The resulting model uses the geological and geophysical data as predictors with the number of earthquakes with magnitude $m_b \geq 4.5$ as target. The rationale for using this method was: (a) to find parameters that correlate with earthquake occurrence and, (b) to generate an automated multivariate rule-based seismicity model using these parameters and their values to predict number of earthquakes with $m_b \geq 4.5$. Because many towns, and villages in Iran have poor resistance against earthquakes a threshold magnitude of $M_C = 4.5$ (i.e. the magnitude of completeness for the earthquake catalog of Iran) has been selected (Zamani and Agh-Atabai 2009, 2011). In the case of events with $m_b \geq M_C$, there was an improvement in the epicentral locations as more instruments were added to the worldwide network of seismological stations (Berberian 1979). On the other hand, because of the poor statistics of the very few large earthquakes, a threshold magnitude of $M_C = 4.5$ makes model validation easier.

Typically, the geological and geophysical variables gathered are not only correlated with each other, but each attribute is also influenced by the other attributes (Zamani and Hashemi 2004; Zamani et al. 2011). The details of exactly how variables such as, magnetic anomalies, gravity anomalies, rock types, fracture systems, earthquakes etc. evolved as a consequence of geodynamic processes is not of our concern in this paper. We wanted to discover the correlations between earthquakes and these variables. In other words, we wanted to identify and extract rules satisfying some minimum confidence threshold and showing the association or coincidence between the predictor variables (splitting variables) and the predicted variable (target variable, i.e. the number of earthquakes with $m_b \geq 4.5$). Decision tree rule extraction model is not intended to investigate the earthquake cycle or the causes of earthquakes in Iran. As the term suggests, data mining using decision tree rule extraction technique has a somewhat more exploratory rather than confirmatory nature. This technique is directed toward searching deeply into characteristics of the large data bases for patterns regardless of cause-and-effect relationships. In data mining, a decision tree described data but not decisions; rather the resulting classification or prediction rules can be an input for decision making. The main objective of this paper is to present a new class of data-driven multivariate rule-based model to create online as well as offline interactive earthquake hazard map that is flexible and readily automated.

2 Seismotectonic setting

The Iranian plateau with its flanking orogens comprises one of the most seismically active areas in the world. This plateau can be characterized by: active faults, recent volcanoes and high surface elevation along the Alpine-Himalaya orogenic belt. The seismic records of Iran are divided into historical document records (pre-1900) and instrumental records (post-1900). Ambraseys and Melville (1982) documented Iran's historical earthquakes. The seismicity studies of Iran from instrumental records was conducted by Wilson (1930), Niazi and Basford (1968), Nowroozi (1971, 1976), Ambraseys and Monifar (1973), Berberian (1973, 1995), Tchalenko (1975), Ambraseys (2001), Engdahl et al. (1998), Engdahl et al. (2006). Reliable earthquake data for Iran exist only for the last few decades because the

locations of earthquakes have been recorded accurately only after the mid 1960s. Historical and instrumental catalogues have shown a spatial correlation between seismicity and seismotectonic sources in Iran. In the last century, many destructive earthquakes occurred in Iran, for example: Silakhor (Ms = 7.4, 1909); Salmas (Ms = 7.4, 1930); Torud (Ms = 6.4, 1953); Lar (Ms = 6.7, 1960); Buyin Zahra (Ms = 7.2, 1962); Dasht-e-Bayaz (Ms = 7.4, 1968); Qir (Ms = 6.9, 1972); Khorgu (Ms = 7, 1977); Tabas (Ms = 7.7, 1978); Qayen (Ms = 7.1, 1979); Rudbar-Manjil (Ms = 7.2, 1990); Birjand (Ms = 7.3, 1997) and Bam (Ms = 6.6, 2003). The seismicity map of Iran (Niazi and Basford 1968) indicates high dispersion and inhomogeneity of the seismic activity in Iran. For example, Earthquakes larger than Ms = 7 have not occurred in the Zagros region; however, shocks of magnitude over Ms = 7 have occurred in Eastern and Central Iran. Seismicity data show that the Zagros fold thrust belt in the southwest, the Kopeh Dagh active fold belt in the northeast, the Alborz thrust belt in the north and the Makran in the south east are the most active areas in Iran. Central and Eastern Iran (minus the Tabas block) are less active. The creation of a possible earthquake hazard map requires the delineation of seismotectonic provinces with high earthquake hazard potential. The seismotectonic provinces of Iran are defined as geographic regions with equal seismic potential and similar geological structures. During the past decades, many researchers have studied and produced tectonic and seismotectonic maps of Iran based on the geological and seismological data, for example: Stöcklin (1968); Berberian (1976, 1977); Nowroozi (1976, 1979); Tavakoli (1996); Tavakoli and Ghafory-Ashtiany (1999); Alavi (1991); Zamani and Hashemi (2004); Zamani et al. (2011), Fig. 1.

3 Method of analysis

One of the ultimate goals of earthquake hazard studies is to understand the distribution of earthquake and earthquake related phenomena in as much detail as possible. Only for the few areas where major faults are identified, this type of study has been made (Rogers et al. 1998). However, this type of study is time consuming and is not possible for every seismogenic regions.

In recent years, machine learning and knowledge discovery techniques have been used for rule extraction in many different fields including business, social sciences, planning, biological sciences, and engineering, among others. These methods include data mining tools such as decision tree, neural network, and rough sets (Fu 1999; Pawlak and Slowinski 1994; Quinlan 1993). The extracted rules reveal trends, patterns, and relationships which might otherwise have remained obscured by the complex patterns of association and massive amount of data.

4 Decision tree

In data mining tools, decision tree is a data driven, non parametric technique without a priori assumptions, which is better suited for non-normal and non-homogeneous data sets. In decision tree theory, a decision tree is a classifier with the structure of a tree with a top-down geometry/hierarchy (Fig. 2) used in statistics, data mining and machine learning. The classification proceeds from top to bottom and has only splitting paths (burst nodes) but no converging paths (sink nodes).This technique is often considered to improve knowledge representation structure by deriving meaningful decision rules and maximizing differences on a dependent variable (Daubie et al. 2002).

The extracted rules are easily interpretable allowing complex relationships to be represented in an intuitive and comprehensible manner. The rules establish a relationship between

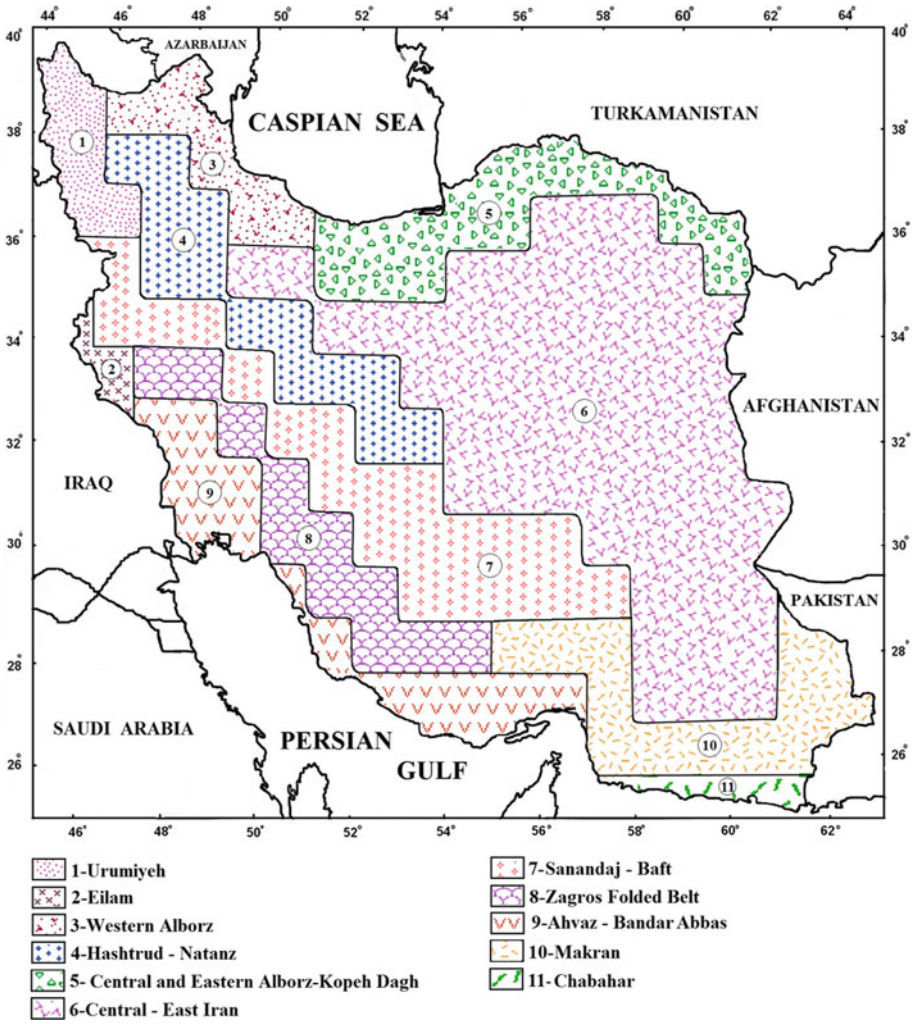

Fig. 1 Automatic integrated self-organized optimum zoning (AISOOZ) map of Iran representing 11 optimum tectonic zones (Zamani et al. 2011)

descriptions of objects by attributes and their assignment to a specific class. This technique eliminates unnecessary or redundant attributes from classification. Decision tree learning produces a directed decision tree as a predictive model. This maps observations about an item leading to conclusions about the item's target value. It describes a tree structure for sequential partitioning of the dataset in order to maximize differences on a dependent variable. In this method, an instance is classified by starting at the root node of the decision tree and testing the attribute specified by this node. The model then moves down the tree branch corresponding to the value of the attribute to some internal node.

This process is then repeated at the node on this branch and so on until a leaf node is reached. This provides the classification of the instance. Each node in the tree specifies a

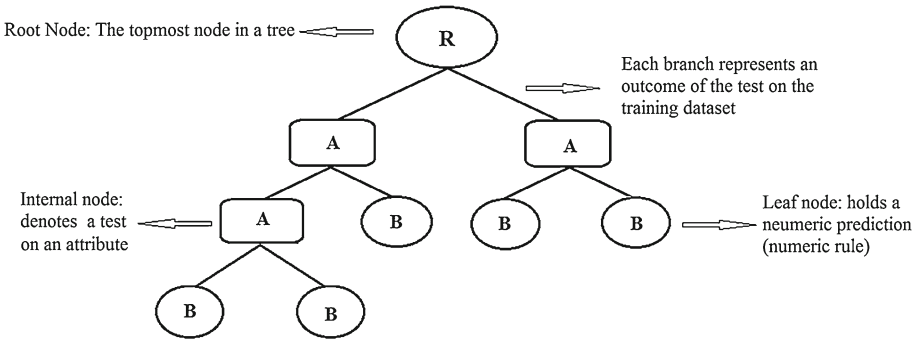

Fig. 2 A typical binary decision tree. (*R*) Root node. (*A*) The internal node. (*B*) The leaf node or terminal node also known as child node. A leaf represents, the predicted value of target attribute given the values of the attributes represented by the path from the root

test of some attribute of the instance. Each node in the tree specifies a test of some branch descending from that node corresponds to one possible value for this attribute.

This paper applies modern data analytical and sorting techniques to develop a useful new type of earthquake hazard map that is flexible and readily automated. This new approach is based on data mining and machine learning technique which uses a decision tree as a predictive model. The model is constructed explicitly or implicitly by inductive learning from a sufficient number of training examples (Mitchell 1997). Classic decision tree learner CART algorithm is used to induce a decision tree model. This rule induction algorithm provides a general framework that can be instantiated in various ways to produce different decision trees (Breiman et al. 1984; Ripley 1996). The underlying assumption of inductive approach is that the trained model is applicable to future, unseen examples, in order to discover unknown patterns. CART analysis is an umbrella term used to refer to both of classification and regression trees analysis procedures (Breiman et al. 1984). For categorical outputs (i.e. classification trees): the leaves of the tree represent a class or group labels. For numeric outputs (i.e. prediction trees): the leaves of the tree predict an average value (i.e. regression trees) or specify a function that can be used to predict the value (i.e. model trees).

CART algorithm offers advantages to other methods of analyzing alternatives. It is inherently non-parametric. That is to say no assumptions are made regarding the underlying distribution of the predictor variables. Thus, CART can handle numerical data that are highly skewed or multi-modal, as well as categorical predictors with either ordinal or non-ordinal structure (Lewis 2000). CART also identifies "splitting" predictor variables based on an exhaustive search of all possible independent variables as splitters. CART handles missing primary splitters (predictor variables) in the data sets by substituting surrogate variables. A primary splitter variable is the best splitter of a node while a surrogate splitter is a splitter that splits in a fashion similar to the primary. It is a relatively automatic machine learning technique. In other words, compared to the complexity of the operations, relatively little input is required from the user. Finally CART decision tree models are easy to interpret even for non-statisticians. The users of CART algorithm aim to classify or predict the values of new examples by feeding them into the root of the tree, and determining which leaf the example flows to. In CART decision tree modelling, a desirable model is one having a relatively small number of directional branches or links, a relatively small number of nodes from which these branches diverge, and high predictive power, in which entities are correctly classified or predicted at the terminal nodes or leaves. In these tree structures, leaves represent

classifications and branches represent conjunction of features or attribute values that lead to these classifications. Each output or terminal value measures the result of a scenario: the sequence of decisions and events on a unique path leading from the root or initial decision node to a specific terminal or leaf node.

5 Data analysis

Earthquakes can be related to the manifestations as well as to the vestiges of long-lived, deep-seated geodynamic processes within the Earth. Meanwhile, almost all surficial features and phenomena of the Earth are fundamental consequences of the interactions of exogenic and endogenic processes (Fowler 2005; Petersen et al. 2011; Turcotte and Schubert 2002). Therefore, identifying earthquake precursors is a difficult task. Despite considerable research efforts by seismologists, scientifically reproducible earthquake predictions can not yet be made (Hough et al. 2009; Pulinets 2006). Conventional techniques fall short of complying with stringent constraints and assumptions to be used in identifying, detecting, and measuring some kind of earthquake precursory phenomena. Furthermore, in seismogenic regions with low seismicity rate, it is difficult to collect statistically significant number of records to derive conclusive prediction. The classical statistical techniques also include a priori assumptions on the data distribution which are difficult to be satisfied for earthquake data sets.

Typically, the geophysical and geological characteristics gathered are not only correlated with each other, but each attribute is also influenced by the other attributes. Thus, in many instances the attributes are interwoven in such a way that when analyzed individually they yield little information about the region under investigation (Zamani et al. 2011). The details of exactly how magnetic anomalies, gravity anomalies, rock types, fracture systems, earthquakes etc. evolved as a consequence of geodynamic processes is not our concern in this paper. We present a novel application of decision tree-based rule extraction method for discovering hidden patterns in these attributes irrespective of contributory causes. In other words, we wanted to identify and extract rules satisfying some minimum confidence threshold and showing the coincidence between independent predictor variables (i.e. geophysical and geological parameters) and dependent predicted variable (i.e. the number of earthquakes with $m_b \geq 4.5$ or target variable). For this purpose, the study area (Iran) is divided into 175 quadrangles, each covering a degree of latitude and longitude (Zamani and Hashemi 2004; Zamani et al. 2011). The quadrangles from west to east are numbered beginning with 1 for the quadrangle between 44°E and 45°E meridians and increasing to 175 for quadrangle between 61°E and 62°E meridians. None of offshore Iran is included in the data set. These quadrangles are used as cases or observations (input samples). Each case has been characterized by 48 variables (attributes) that seem to characterize the intensity and degree of contrast between tectonic and seismotectonic structures in Iran (Table 1).

As mentioned earlier, CART software package for decision tree building is used with our proposed data set presented in Table 1. The rational for using CART was: (a) to find the best independent or splitter attributes (i.e. those that are most closely related to the number of earthquakes with $m_b \geq 4.5$); and (b) to extract rules from the resultant decision tree model using these attributes and their values, which allow us to predict the number of earthquakes with $m_b \geq 4.5$. The inductive learning of decision tree in this paper involves the processing of a relatively large number of geophysical and geological characteristics or attributes. The rules based on the correlates that best predict the number of earthquakes with $m_b \geq 4.5$. They are quite useful for earthquake hazard analysis, particularly in regions where the earthquake cycle is relatively slow or recording stations are inadequate to collect significant

Table 1 Attributes used for constructing the multivariate rule-based seismicity map (MRBSM), measured within 1° quadrangles

No.	Attributes	No.	Attributes
1	a- value in the Gutenberg–Richter's formula, **AVGRF**	25	Minimum gravity anomaly (mgal), **MIGRV**
2	b- value in the Gutenberg–Richter's formula, **BVGRF**	26	Range of free air anomaly (mgal), **RAFRA**
3	Maximum earthquake magnitude (m_b), **MXEMG**	27	Average free air r anomaly (mgal), **AVFRA**
4	Number of earthquakes greater than $m_b \geq 4.5$, **NEGMB**	28	Maximum free air anomaly (mgal), **MXFRA**
5	Maximum seismic energy released (j), **MXSER**	29	Minimum free air anomaly (mgal), **MIFRA**
6	Range of isostatic anomaly (mgal), **RA ISO**	30	Range of magnetic intensity (gamma), **RAMGI**
7	Average isostatic anomaly (mgal), **AVISO**	31	Average magnetic intensity (gamma), **AVMGI**
8	Maximum isostatic anomaly (mgal), **MXISO**	32	Maximum magnetic intensity (gamma), **MXMGI**
9	Minimum isostatic anomaly (mgal), **MIISO**	33	Minimum magnetic intensity (gamma), **MIMGI**
10	Range of regional Bouger anomaly (mgal), **RAEGB**	34	Average Moho depth (km), **AVMOD**
11	Average regional Bouger anomaly (mgal), **AVREG**	35	Range of elevation (m), **RAELV**
12	Maximum regional Bouger anomaly (mgal), **MXREG**	36	Average elevation (m), **AVELV**
13	Minimum regional Bouger anomaly (mgal), **MIREG**	37	Maximum elevation (m), **MXELV**
14	Range of residual Bouger anomaly (mgal), **RARES**	38	Minimum elevation (m), **MIELV**
15	Average residual Bouger anomaly (mgal), **AVRES**	39	Relative area of surface unconsolidated sediment cover(%), **RAUNR**
16	Maximum residual Bouger anomaly (mgal), **MXRES**	40	Relative area of surface sedimentary rocks (%), **RASER**
17	Minimum residual Bouger anomaly (mgal), **MIRES**	41	Relative area of surface metamorphic rocks (%), **RAMER**
18	Range of Bouger anomaly (mgal), **RABUG**	42	Relative area of surface igneous rocks (%), **RAIGR**
19	Average Bouger anomaly (mgal), **AVBUG**	43	Relative area of surface ophiolitic rocks (%), **RAOPR**
20	Maximum Bouger anomaly (mgal), **MXBUG**	44	Relative area of surface Cenozoic rocks (%), **RACER**
21	Minimum Bouger anomaly (mgal), **MIBUG**	45	Relative area of surface Mesozoic rocks (%), **RAMER**
22	Range of gravity anomaly (mgal), **RAGRV**	46	Relative area of surface Paleozoic rocks (%), **RAPAR**
23	Average gravity anomaly (mgal), **AVGRV**	47	Relative area of surface Proterozoic rocks (%), **RAPTR**
24	Maximum gravity anomaly (mgal), **MXGRV**	48	Fault length density (km^{-1}), **FLTLD**

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 2010 (Engdahl et al. 2006; Gutenberg and Richter 1954; ISC 2012; NEIC 2012). Geophysical data have been taken from Dehghani and Makris (1983), total magnetic intensity maps of Iran (Yousefi 1989), Seismicity and fault map of Iran (Mohajer-Ashjai and Nabavi 1982) and digital data from Geological Survey of Iran

number of records. For this propose, the database of attributes measurement represented in Table 1 is compiled for the 175 quadrangular observation sites of 1○ area. As the input variables were continuous, computer program for CART analysis was run in regression—tree mode. Tree building process start at the root, which includes all observations in the full learning database (i.e. the source set). Starting with this node, the CART computer algorithm finds the best possible independent variable to split the root node into two child nodes, based on an exhaustive search of all possibilities of splitter variables. These child nodes are then split and so forth. In each node the optimum decision rule (i.e. the best splitter variable) is found based some impurity measure for the two child nodes. For regression tree (i.e. prediction tree) analysis the sum of squared errors (SSE), the so called sum of squared deviations or variance is automatically applied. At each node the recursive procedure may stop splitting the node further when the variation or SSE between the predicted model and observed values is minimized. A small SSE indicates a tight fit of the model to the observed data. The procedure results in the selection of the independent variable that produces the greatest "separation" in the target variable (i.e. the number of earthquakes with $m_b \geq 4.5$). Splitting procedure stops when there is only one data-point left in each node, or when each node has only the same predicted values. At this point a "maximal" tree has been constructed, which probably overfits, because it represents all idiosyncrasies of the learning data set. For regression tree (i.e. prediction tree), a "maximal" tree is achieved such that means between nodes vary as much as possible and standard deviation or variance (i.e. dispersion) within each node is as low as possible. The maximal tree has the maximum number of levels of tree growth beneath the root node, where the maximum tree depth is reached. The maximum regression tree is unable to split the data without violating a condition that some parent node has observations of different classes. In search for useful patterns in data set it is essential to avoid the trap of overfitting or finding patterns that apply only to the training data set. CART' s embedded test disciplines ensure that the patterns found will hold up when applied to new data set. Further, the testing and selection of the optimally-sized tree are an integral part of the CART algorithm.

The optimal regression tree was found by tenfold cross-validation which is an industry standard for many applications (Oates and Jensen 1997). Hence, CART uses tenfold cross-validation for maximum regression tree optimizations. This was done by dividing our learning data set into ten subsets with an equal distribution for the dependent (target) variable (i.e. the number of earthquakes with $m_b \geq 4.5$). A maximal regression tree was grown from 90 % of the subsets, with 10 % of the data set reserved for assessing the SSE. This process was repeated ten times with the learning data, each time reserving a different 10 % of the data for SSE assessment.

Error rates from the regression trees were combined to yield estimated error rates for the nodes in the maximum tree (Steinberg and Colla 1997). But employing cross-validation method itself for all possible regression trees with different sizes is also not feasible due to computational constraints. Therefore, CART algorithm dose not check all subtrees of the maximum regression tree but only special key subtrees. For this purpose, some new measure the so called cost-complexity index is automatically applied in the CART algorithm (Breiman et al. 1984). This index controls the size of the resulting regression tree which can be estimated by the number of terminal nodes. The idea is that the maximum regression tree will get a penalty for its big size; on the other hand, it is more accurate for predicting. Small regression trees will get relatively lower penalty, for their size but their predicting abilities are limited. Optimization procedure based on such trade-off criterion could determine the optimally-sized regression tree with the help of cost-complexity function and cross-validation (Breiman et al. 1984). Our maximum regression tree was grown, and then pruned back to

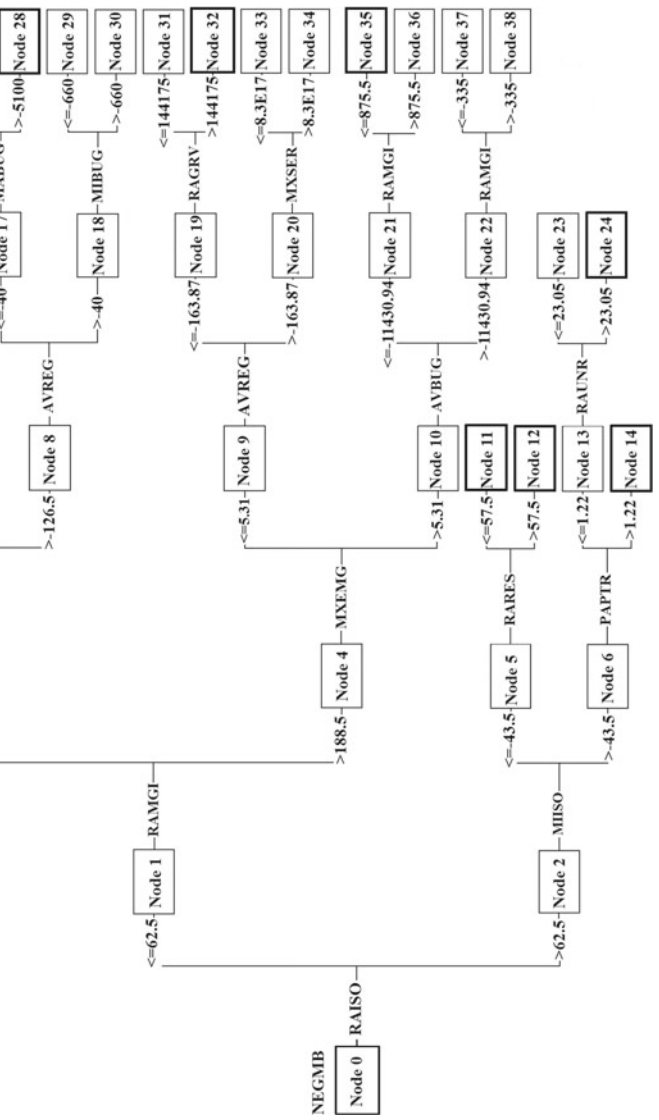

Fig. 3 CART binary decision tree for predicting earthquakes occurrence with $m_b \geq 4.5$. Nodes with gain index values greater than 100 % were chosen as influential predictors. Nodes with gain index values greater than 100 % are highlighted in figure

obtain the optimal size tree by determining the lowest SSE. In CART algorithm, the automatic setting limits the maximum tree growth to five levels beneath the root node.

6 Results and discussion

Many investigators have suggested that seismicity in Iran is related to geographical position, local geology and tectonics. Such studies often lack adequate accuracy since the earthquake cycle is slow or recording stations are inadequate (Lomnitz 1994; Zamani and Agh-Atabai 2009, 2011). In this work, a new and updated catalogue of geological and geophysical data of Iran has been used as training data to model the occurrence of earthquakes. Statistically significant rules associated with the number of earthquakes with $m_b \geq 4.5$ (the target variable) are found (Fig. 3).

The target variable information is provided by the gain index values shown in Table 2. This index is a useful tool for measuring the value of our decision tree predictive model. The index value is basically an indication of how far the observed target category percentage for that node differs from the expected target category percentage in the root node before the effects of any of the independent variables are considered. The gain index percentage tells

Table 2 The gain summery provides statistics for all terminal nodes in the tree

Rule no.	Node number	Number of observations in node	Node percentage	Predicted values	Gain index value (%)
1	12	6	3.4	67	436.6
2	26	3	1.7	54	349.7
3	14	2	1.1	48	310.1
4	11	6	3.4	40	262.7
5	25	7	4	38	247.1
6	28	2	1.1	28	181.6
7	24	4	2.3	22	144.4
8	15	3	1.7	21	140.2
9	32	2	1.1	18	118.3
10	35	19	10.9	17	110.5
11	31	2	1.1	14	94.1
12	37	8	4.6	14	92.7
13	27	7	4	13	85.5
14	36	13	7.4	12	75.3
15	23	5	2.9	9	59.1
16	38	26	14.9	9	57.7
17	34	41	23.4	6	36.9
18	29	3	1.7	5	31.3
19	33	14	8	3	19.9
20	30	2	1.1	1	9.7

Node number: the number of node in Fig. 3. Number of observations in the node: The total number of samples at that node. Node percentage: The percentage of all samples in the dataset that fall into this node. Predicted values: The predicted target for each node

Table 3 The ten most reliable decision tree-based rules (i.e. IF-THEN rules) with gain index value of more than 100 %

Rule no.	Node no.	IF	THEN
1	12	RAISO(6) > 62.5 and MIISO(9) ≤ −43.5 and RARES(14) > 57.5	NEGMB(4) = 67
2	26	RAISO(6) ≤ 62.5 and RAMGI(30) ≤ 188.5 and MIREG(13) ≤ −126.5 and AVISO(7) > −23.2 and MIBUG(21) > −11835	NEGMB(4) = 54
3	14	RAISO(6) > 62.5 and MIISO(9) > −43.5 and RAPTR(47) > 1.2	NEGMB(4) = 48
4	11	RAISO(6) > 62.5 and MIISO(9) ≤ −43.5 and RARES(14) ≤ 57.5	NEGMB(4) = 40
5	25	RAISO(6) ≤ 62.5 and RAMGI(30) ≤ 188.5 and MIREG(13) ≤ −126.5 and AVISO(7) > −23.2 and MIBUG(21) ≤ −11835	NEGMB(4) = 38
6	28	RAISO(6) ≤ 62.5 and RAMGI(30) ≤ 188.5 and MIREG(13) > −126.5 and AVREG(11) ≤ −40 and MXBUG(20) > −5100	NEGMB(4) = 28
7	24	RAISO(6) > 62.5 and MIISO(9) > −43.5 and RAPTR(47) ≤ 1.2 and RAUNR(39) > 23.05	NEGMB(4) = 22
8	15	RAISO(6) ≤ 62.5 and RAMGI(30) ≤ 188.5 and MIREG(13) ≤ −126.5 and AVISO(7) ≤ −23.2	NEGMB(4) = 21
9	32	RAISO(6) ≤ 62.5 and RAMGI(30) > 188.5 and MXEMG (3) ≤ 5.3 and AVREG(11) ≤ −163.9 and RAGRV(22) > 144175	NEGMB(4) = 18
10	35	RAISO(6) ≤ 62.5 and RAMGI(30) > 188.5 and MXEMG(3) > 5.3 and AVBUG(19) ≤ −11430.9 and RAMGI(30) ≤ 875.5	NEGMB(4) = 17

The number listed after each set of initials in Table 3 is its attribute in Table 1

us how much greater the proportion of a given target at each node differs from the overall proportion.

An index value greater than 100 means the percentage of cases in the target category in the node exceeds the percentage in the root node. Nodes with gain index values greater than 100 % indicate that a better chance exists of accurate prediction by selecting records from these nodes instead of random selection from the entire sample. The index values in this paper show that node 12 has the highest possible rate for the entire data, with a value of 436 %. This node is thus almost 4.4 times more likely to get a hit with these records than using a random selection. The gain index values show that of the 20 nodes, 10 have index values greater than 100 %. The rules of the top ten nodes are depicted in Table 3.

The CART methodology sorts the predictor variables in decreasing order of importance. Interestingly, the constructed decision tree model (Fig. 3) indicates that the isostatic anomaly is a very important parameter in earthquake prediction. Other important factors in decreasing order of importance include: magnetic anomaly, Bouger anomaly, and gravity anomaly respectively. Such results support our previous researches (Zamani and Hashemi 2000; Zamani and Farahi Ghasre-Aboonasr 2011) which showed that there is a strong correla-

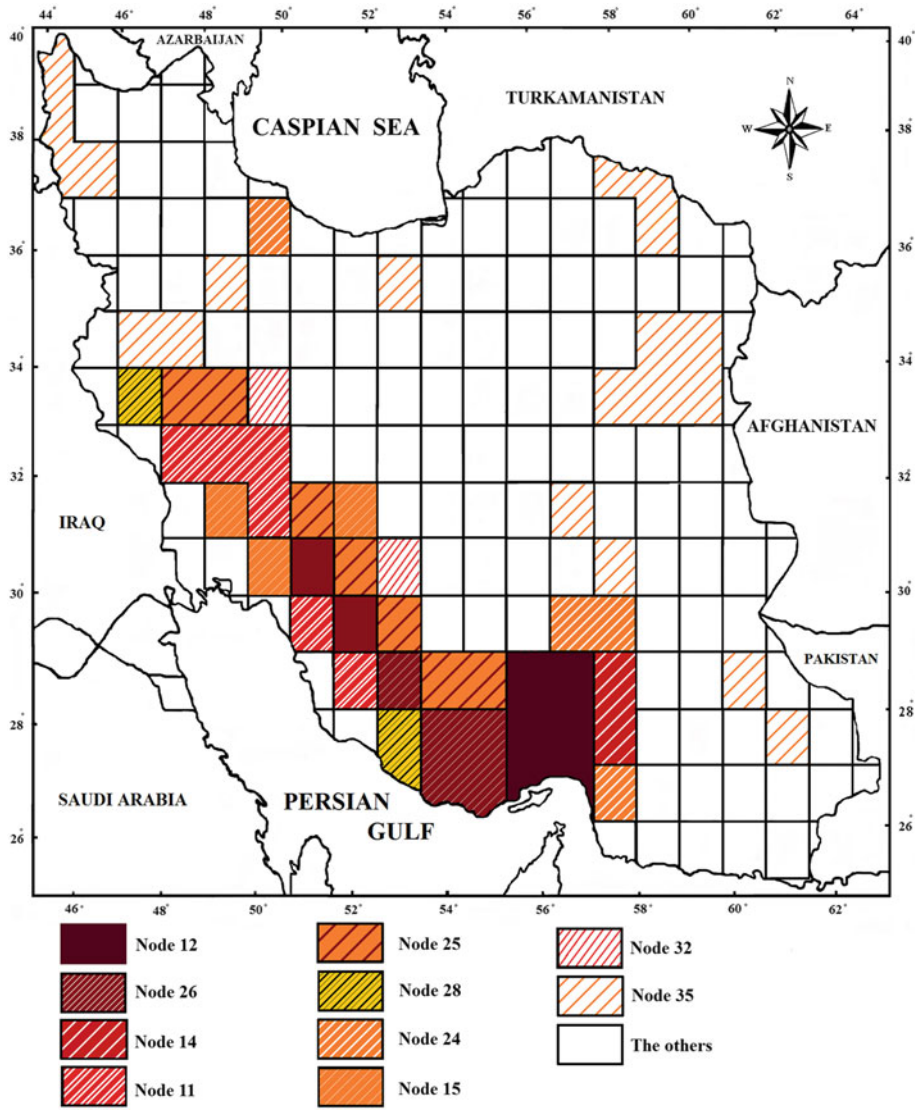

Fig. 4 The multivariate rule-based seismicity map (MRBSM) of Iran. Indexes are shown in decreasing order of importance for predicting future earthquakes. For example, node 12 has the highest hazard for future earthquakes with $m_b \geq 4.5$. None of offshore Iran and island is included in the data set

tion between seismicity and gravity anomalies in Iran. It seems that, isostatic and Bouger anomalies caused by the regional variations in lithospheric thickness and/or in density, affect gravitational stability and thereby the differential stresses responsible for earthquakes.The results further suggest that despite abundant use in earthquake studies, of Gutenborg-Richter a- and b-values, these parameters have low correlations with the occurrences of earthquakes with $m_b \geq 4.5$. This could be due to the fact that in seismogenic regions where the earthquake cycle is slow and/or seismic station coverage is inadequate, it is difficult to collect statistically significant number of records to determine a- and b-values accurately.

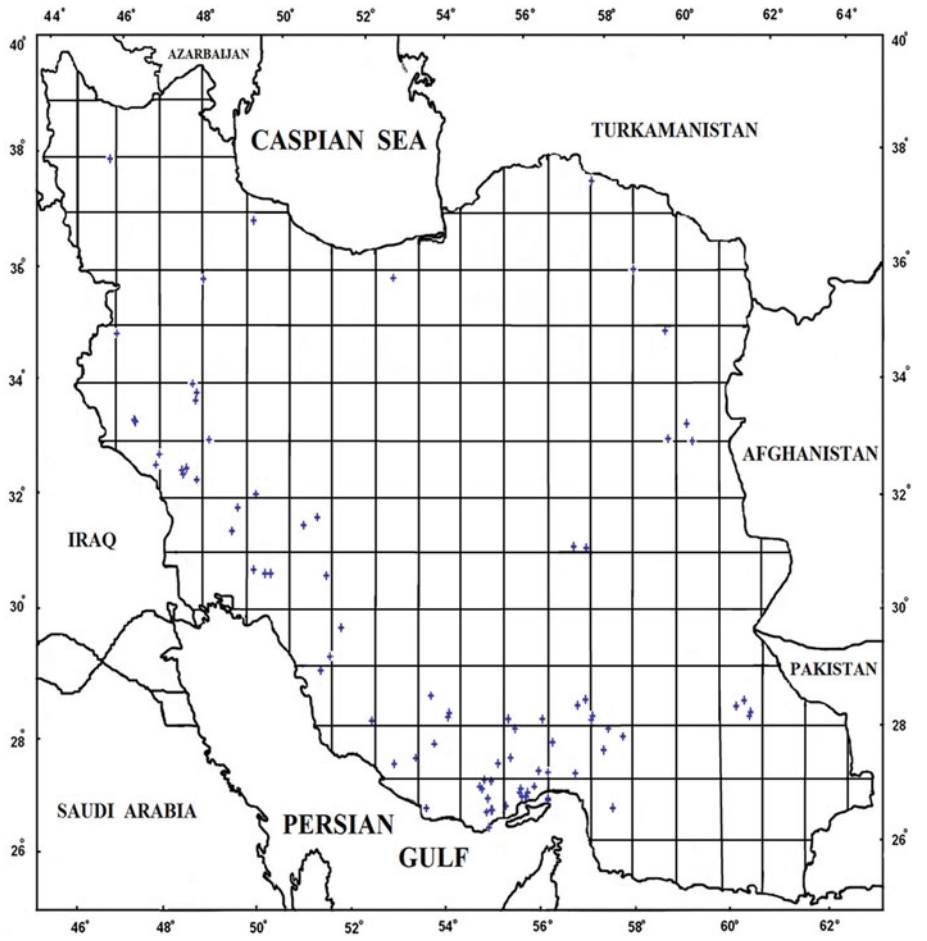

Fig. 5 Seismicity map of Iran based on 2008 to end 2010 earthquakes with $m_b \geq 4.5$ (NEIC 2012). This map correlates well with multivariate rule-based seismicity map (MRBSM) of Iran (Fig. 4). This indicates that the rules applied have high accuracy

In this work, Multivariate Rule-Based Seismicity Map (MRBSM) is defined as the map of regions with a high hazard of future earthquakes with $m_b \geq 4.5$. Use of the above-mentioned model produced the MRBSM of Iran (Fig. 4)

For further evaluating the performance of the model, it is applied to test set to predict the 2008–2010 earthquake records of previously unseen data (Fig. 5).

When comparing Figs. 4 and 5, one can see that with the exception of a few cases, virtually the entire earthquake records from the beginning of 2008 to the end of 2010 closely matched the MRBSM of Iran. The map indicates that Bandar Abbas in the South of Iran, parts of the Zagros simply folded belt with its NW-SE trend, the Oman line (a limited area in southern Iran) and the northern portion of the Lut block in eastern Iran are regions of high hazard from future earthquakes with $m_b \geq 4.5$.

The result indicates that the novel approach introduced in this paper is a reliable method for seismic hazard assessment in Iran.This conclusion is worthy of further study to find out if this approach contributes to the earthquake hazard assessment in other seismogenic regions.

7 Conclusions

In this paper, a novel approach based on the decision tree rule-extraction technique for earthquake hazard assessment is presented. For this purpose a new and updated catalogue of geological and geophysical data from Iran are used to predict (by rule extraction) the number of future earthquakes with $m_b \geq 4.5$. The rules extracted from among the attribute were significant statistically to assure the mapped patterns are not random and actually relate to earthquake locations. The CART algorithm was used in a regression-tree mode for prediction and for rule extraction. The rules extracted were used to produce a future earthquake hazard map (MRBSM) for Iran (Fig. 4). This map shows onshore regions with high hazard for future earthquakes occurring with $m_b \geq 4.5$. These regions are the Bandar Abbas area in the South of Iran, the Zagros simply folded belt with its NW-SE trend, the Oman line in southern Iran and the northern portion of Lut block in eastern Iran. The analysis also shows that the isostatic anomaly correlates best with these earthquakes. Other important factors in decreasing order of importance are: magnetic intensity, regional Bouger anomaly, Bouger anomaly and gravity anomaly, respectively. The results further suggest that despite the widespread use in earthquake analysis of a-and b-values from the Gutenberg-Richter formula, these parameters have low correlation with the earthquake occurrence in Iran.

The results presented in this paper indicate that the novel approach based on decision tree rule extraction model is a reliable method to assess earthquake hazard in Iran. The conclusion is worthy of further study to find out if this approach contributes to the earthquake hazard analysis in other seismogenic regions.

Acknowledgments We are grateful to two anonymous reviewers for their detailed reviewing of the manuscript along with their helpful suggestions. The assistance of the Editor-in-Chief is also appreciated. This study was supported by the Centre of Excellence for Environmental Geohazards and the Research Council of Shiraz University.

References

Alavi M (1991) 1/5,000,000 sheet, tectonic map of middle east. Geol Suev Iran

Ambraseys NN, Monifar A (1973) The seismicity of Iran. The Silkhor, Lurestan, Erthquake of 23rd January 1909. Ann Geophys 4: 659–678

Ambraseys NN, Melville CP (1982) A history of persian earthquakes. Cambridge Univ Press, Cambridge 219

Ambraseys NN (2001) Reassessment of earthquakes, 1900–1999, in the Eastern Mediterranean and the Middle East. Geophys J Int 45(2): 471–485

Ashtari Jafari M (2010) Statistical prediction of the next great earthquake around Tehran, Iran. Geodyn 49: 14–18

Berberian M (1973) 1/2500000 sheet, the seismicity of Iran. Preliminary map of epicenters and focal depth. Geol Suev Iran

Berberian M (1976) Contribution to the seismotectonics of Iran (part II). Geol Suev Iran, Report no 39, Tehran, pp 518

Berberian M (1977) Contribution to the seismotectonics of Iran (part III). Geol Suev Iran, Report no 40, Tehran, pp 300

Berberian M (1979) Discussion of the paper A. A. Nowroozi, 1976 seismotectonic province of Iran. Bull Seismol Soc Am 69:293–297

Berberian M (1995) Natural hazards and the first earthquake catalogue of Iran. IISEE, Iran 603

Berg JW, Gaskell R, Rinehart V (1964) Earthquake energy release and isostasy. Bull Seismol Soc Am 54(2): 777–784

Bonini M, Corit G, Sokoutis D, Vannucci G, Gasperini P, Cloetingh S (2003) Insight from scaled analogue modeling into the seismotectonics of the Iranian region. Tectonophys 376: 149–157

Bouchon M (1973) Effect of topography on surface motion. Bull Seismol Soc Am 63: 615–632

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Inc, Monterey
Caputo M, Milana G, Rayhorn J (1984) Topography and its isostatic compensation as a cause of seismicity of the Apennines. Tectonophys 102: 333–342
Caputo M, Manzetti V, Nicelli R (1985) Topography and its isostatic compensation as a cause of seismicity; a revision. Tectonophys 111: 25–39
Chen YT, Liu KR, Zheng JH, Song SH, Liu RF, Lu HY, Gu FY (2002) A review of the studies on the relationship between local gravity field changes and earthquakes. In: Sun S (ed) Advances in pure and applied geophysics. Meteorology Press, Beijing, pp 40–47 (in Chinese)
Daubie M, Levecq P, Meskens N (2002) A comparison of rough sets and recursive partitioning induction approaches: an application to commercial loans. Int Trans Oper Res 9: 681–694
Davis LL, West LR (1973) Observed effects of topography on ground motion. Bull Seismol Soc Am 63: 283–298
Dehghani GA, Makris J (1983) The gravity field and crustal structure of Iran. In: Geodynamic Project (Geotraverse) in Iran. Geol Suev Iran, pp 51–68
Dmeroski S (2002) Applications of KDD methods in environmental sciences. In: Kloesgen W, Zytkow J (eds) Handbook of data mining and knowledge discovery. Oxford Univ Press, Oxford
Engdahl ER, Vander Hilst RD, Buland RP (1998) Global teleseismic earthquake relocation with improved travel times and procedures for depth determination. Bull Seismol Soc Am 88: 722–743
Engdahl ER, Jackson JA, Myers SC, Bergman E A, Priestley K (2006) Relocation and assessment of seismicity in the Iran region. Geophys J Int 167: 761–778
Fowler CMR (2005) The solid earth: an introduction to global geophysics. Cambridge Univ Press, Cambridge
Fu LM (1999) Knowledge discovery based on neural networks. Commun ACM 42(11): 47–50
Geli L, Bard PY, Jullien BA (1988) The effect of topography ground motion: a review and new results. Bull Seismol Soc Am 78(1): 42–63
Geological Survey of Iran (2004) 1/5,000,000 sheet, geological map of Iran. Ministry of Industries and Mines
Gutenberg B, Richter CF (1954) Seismicity of the earth and its associate phenomena. Princeton University Press, Princeton 310
Hough SE, Bhat I, Bilham R (2009) On shaky ground- megaquakes in Kashmir. Am Sci 97(1): 42–49
Iftikhar US, Toshinori M (2009) Application of rough set and decision tree for characterization of premonitory factors of low seismic activity. Exp Syst Appl 36: 102–110
ISC (2012) International seismological centre. Newbury, Berkshire
Johnston MJS (1997) Review of electric and magnetic fields accompanying seismic and volcanic activity. Surv Geophys 18: 441–475
Karakaisis GF (1994) Long-term earthquake prediction in Iran based on the time and magnitude predictable model. Phys Earth Planet Inter 83: 129–145
Lewis RJ (2000) An introduction to classification and regression tree (CART) analysis. Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance
Li ZX, Li H (2009) Earthquake-Related gravity field changes at Beijing-Tangshan gravimetric network during 1987–1998. Stud Geophys Geod 53: 185–197
Lomnitz C (1994) Fundamental of earthquake prediction. Wiley, New York 326
Mitchell T (1997) Machine learning. McGraw-Hill, New York 400
Mohajer-Ashjai A, Nabavi MS (1982) 1/2,500,000 sheet, seismicity and fault map of Iran. AEOI
NEIC (2012) National earthquake information center. Colorado, USA
Niazi M, Basford JR (1968) Seismicity of Iranian Plateau and Hindu Kush region. Bull Seismol Soc Am 58: 1843–1861
Nowroozi A (1971) Seismotectonics of the Persian Plateau, eastern Turkey, Caucasus, and Hindu-Kush regions. Bull Seismol Soc Am 61: 317–341
Nowroozi A (1976) Seismotectonic provinces of Iran. Bull Seismol Soc Am 66: 1249–1276
Nowroozi A (1979) Reply to M, Berberian comparison between instrumental and macroseismic epicenter. Bull Seismol Soc Am 69: 641–649
Oates T, Jensen D (1997) The effects of training set size on tree size. In: 14th international conference on machine learning, pp 254–262
Pawlak Z, Slowinski R (1994) Decision analysis using rough sets. Int Trans Oper Res 1: 104–107
Petersen JF, Sack D, Gable RE (2011) Physical geography. Thomson Brooks/Cole, Belmont 646
Pulinets SA (2006) Space technologies for short-term earthquake warning. Adv Space Res 37: 643–652
Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufman, San Mateo
Ripley BD (1996) Pattern Recognition and neural networks. Cambridge Univ Press, Cambridge 403
Rogers GC, Cassidy JF, Weichert DH (1998) Variation in earthquake ground motion on the Fraser Delta from strong motion seismograph records. In: Clagve JJ, Luternauer JC, Mosher DC (eds) Recent geological,

geophysical, geotechnical, and geochemical research. Fraser River Delta, British, Columbia. Geologic Surv Canada Bull 525:195–210
Shoja-Taheri J, Niazi M (1981) Seismicity of Iranian Plateau and bordering regions. Bull Seismol Soc Am 71: 477–489
Standart G, Penaloza M, Zong Z (2010) Use of data mining techniques in the discovery of spatial and temporal earthquake relationship. Proceedings of midwest instruments computation symposium
Steinberg D, Colla P (1997) Classification and regression trees. Salford Systems, San Diego
Stöcklin J (1968) Structural history and tectonics of Iran: a review. Am Assoc Petrol Geol Bull 52: 1229–1258
Tavakoli B (1996) Major seismetectonic provinces of Iran. unpublished map. IIEES, inter doc (in Persian)
Tavakoli B, Ghafory-Ashtiany M (1999) Seismic hazard assessment of Iran. Annali DI Geofisica 42: 1013–1021
Tchalenko JS (1975) Seismicity and structure of the North Tehran fault. Tectonophys 29: 411–420
Turcotte D, Schubert G (2002) Geodynamics-applications of continuum physics to geological problems. Cambridge University Press, Cambridge 450
Wilson AT (1930) Earthquakes in Persia. Bull Sch Orient Stud Lond 6: 103–131
Yousefi E (1989) 1/250,000 sheets, total magnetic intensity maps of Iran. Geol Suev Iran
Zamani A, Hashemi N (2000) A comparison between seismicity, topographic relief, and gravity anomalies of Iranian Plateau. Tectonophys 327: 25–36
Zamani A, Hashemi N (2004) Computer-based self-organized tectonic zoning: a tentative pattern recognition for Iran. Comput Geosci 30: 705–718
Zamani A, Agh-Atabai M (2009) Temporal characteristics of seismicity in the Alborz and Zagros region of Iran, using a multifractal approach. J Geodyn 47: 271–279
Zamani A, Agh-Atabai M (2011) Multifractal analysis of the spatial distribution of earthquake epicenters in Zagros and Alborz-Kopeh Dagh region of Iran. IJST A1: 39–51
Zamani A, Farahi Ghasre-Aboonasr S (2011) The significance of parameters used for selforganized tectonic zoning of Iran. Iran Geosci J 81:165–170 (in Persian)
Zamani A, Khalili M, Gerami A (2011) Computer-based self-organized zoning revisited: scientific criterion for determining the optimum number of zones. Tectonophys 510: 207–216
Zmazek B, Todorovski L, Dzeroski S, Vaupotic J, Kobal I (2003) Application of decision trees to the analysis of soil radon data for earthquake prediction. Appl Radi Isot 58: 697–706