Seismic source modeling by clustering earthquakes and predicting earthquake magnitudes

Mahdi Hashemi*, Hassan A. Karimi

Geoinformatics Laboratory, School of Information Sciences, University of Pittsburgh,
135 North Bellefield Avenue, Pittsburgh, Pennsylvania 15260, U.S.
m.hashemi1987@gmail.com, hkarimi@pitt.edu

Abstract. Seismic sources are currently generated manually by experts, a process which is not efficient as the size of historical earthquake databases is growing. However, large historical earthquake databases provide an opportunity to generate seismic sources through data mining techniques. In this paper, we propose hierarchical clustering of historical earthquakes for generating seismic sources automatically. To evaluate the effectiveness of clustering in producing homogenous seismic sources, we compare the accuracy of earthquake magnitude prediction models before and after clustering. Three prediction models are experimented: decision tree, SVM, and kNN. The results show that: (1) the clustering approach leads to improved accuracy of prediction models; (2) the most accurate prediction model and the most homogenous seismic sources are achieved when earthquakes are clustered based on their non-spatial attributes; and (3) among the three prediction models experimented in this work, decision tree is the most accurate one.

Keywords: Clustering; Prediction; Seismic source; Earthquake, Big data.

1 Introduction

The study of earthquake ground motions and associated hazards and risks play an important role in sustainable development especially in earthquake-prone areas such as southwestern United States [1; 2]. Reliable evaluation of seismic hazards and risks is a foundation for all earthquake mitigation plans, upon which decision makers can prepare for earthquakes in an optimal way. The first step in any seismic hazard analysis is earthquake source modeling [3; 4]. A single earthquake source is supposed to be uniform in terms of earthquake potential, i.e., the chance of an earthquake of a given magnitude occurring is the same throughout the source. Sources may be linear or areal [4] and are usually used to generate hazard maps and estimate the probability of earthquakes of different magnitudes [5]. Large collections of historical earthquakes have made it possible to construct these sources more efficiently. Seismologists usually determine the boundary of seismic sources manually based on historical earthquakes and tectonic features [6; 7; 8] with no standard or automatic method in place. However, as the size of historical earthquake databases grows, the manual delineation of source boundaries becomes more cumbersome and less accurate. This calls for development of approaches to automate the same process.

Anderson and Nanjo [2] clustered earthquakes based on their distance in space and time and proposed an optimal distance and time interval, obtained experimentally, for clustering earthquakes. Zmazek et al. [9] used a decision tree to predict the radon concentration in soil based on environmental variables. They found that the accuracy of their prediction model changes during seismically active periods comparing with seismically inactive periods. They proposed to predict the time of earthquakes based on this observation. Hashemi and Alesheikh [10] used spatial data mining techniques and indices to reveal the characteristics of earthquakes. They clustered earthquakes around a fault in one class and showed that the earthquake magnitudes in each class are neither spatially correlated nor have any spatial trend, though the earthquakes themselves are strongly clustered at multiple distances. They suggested, as future research, developing prediction models of earthquake characteristics.

The work in this paper is focused on developing a methodology for generating areal seismic sources based on historical earthquakes. Different from previous approaches, the proposed methodology benefits from hierarchical clustering technique [11; 12; 13] and is for the purpose of automating the process. Faults, tectonic features and linear sources are not considered in this work. Three clustering approaches are explored:

a) hierarchical clustering only based on non-spatial attributes,
b) hierarchical clustering only based on location, and
c) hierarchical clustering based on all attributes.

The purpose of clustering is to categorize similar events together. When events are earthquakes, this process coincides with the purpose of seismic source modeling. Thus, assuming similar earthquakes are clustered correctly, one should be able to develop more accurate prediction models in each cluster than without clustering. The proposed prediction model in this work aims to predict the magnitude of an earthquake based on its other characteristics. A different prediction model is required for each cluster. Assuming the first clustering approach (a above) results in n clusters, there should be n prediction models, one for each cluster. Consequently if the second and third clustering approaches (b and c above) result in m and k clusters, respectively, there should be m prediction models for the second one and k prediction models for the third one. Decision tree, SVM and kNN [11; 12; 13] are three different prediction models experimented in this work, resulting in a total of $3\times n\times m\times k$ different prediction models. These prediction models are evaluated using 10-fold cross validation. The accuracy of a prediction model not only reveals the strength and suitability of the applied prediction model (decision tree, SVM or kNN), but also demonstrates the effectiveness of clustering in producing homogenous seismic sources. Thus, by comparing and analyzing the evaluation results, suggestions are made at the end of this article regarding appropriate clustering approaches and prediction models for earthquakes. Fig. 1 shows the process of clustering earthquakes and predicting earthquake magnitudes used in this work.

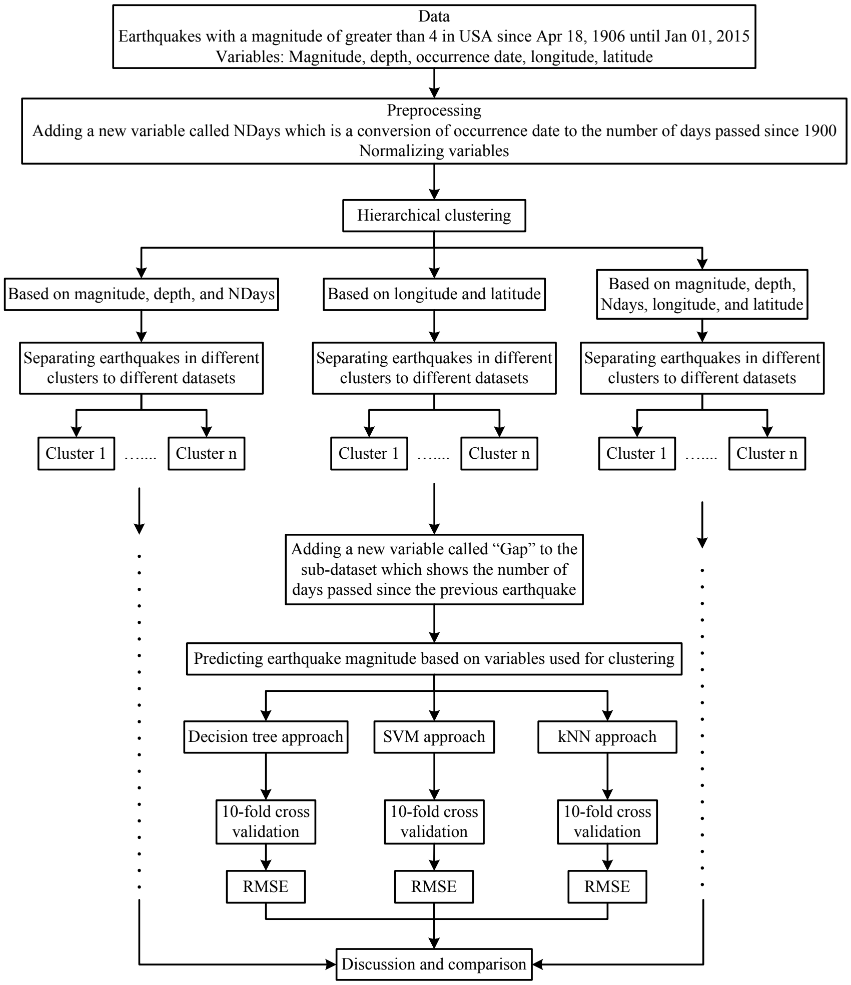
Data
Earthquakes with a magnitude of greater than 4 in USA since Apr 18, 1906 until Jan 01, 2015
Variables: Magnitude, depth, occurrence date, longitude, latitude
Preprocessing
Adding a new variable called NDays which is a conversion of occurrence date to the number of days passed since 1900
Normalizing variables
Hierarchical clustering
Based on magnitude, depth, and NDays
Based on longitude and latitude
Based on magnitude, depth, Ndays, longitude, and latitude
Separating earthquakes in different clusters to different datasets
Separating earthquakes in different clusters to different datasets
Separating earthquakes in different clusters to different datasets
Cluster 1
Cluster n
Cluster 1
Cluster n
Cluster 1
Cluster n
Adding a new variable called "Gap" to the sub-dataset which shows the number of days passed since the previous earthquake
Predicting earthquake magnitude based on variables used for clustering
Decision tree approach
SVM approach
kNN approach
10-fold cross validation
10-fold cross validation
10-fold cross validation
RMSE
RMSE
RMSE
Discussion and comparison

Fig. 1. Process of clustering earthquakes and predicting their magnitudes.

2 Data

Earthquakes with a magnitude of greater than 4 in the United States between April 18, 1906 and January 1, 2015 were downloaded from the United States Geological Survey (USGS) website [14]. This is the longest time range available in the database at the time of writing this article in January 2015. The dataset contains 5,368 earthquakes. Table 1 shows the description of each attribute in the dataset [15].

Table 1. Available attributes for earthquakes.

Variable	Description
Longitude	Decimal degrees longitude. Negative values for western longitudes.
Latitude	Decimal degrees latitude. Negative values for southern latitudes.
Magnitude	The magnitude for the event.
Depth	Depth of the event in kilometers.

Although the occurrence dates of earthquakes are available in the original dataset, they are formatted here as the number of days passed since 1900 and called NDays in the dataset. Since all earthquakes in the dataset have occurred after 1900, all values for this variable are positive.

An important step before clustering data points or developing prediction models is normalizing variables. To normalize a variable (e.g., Magnitude in the dataset), the values are transformed to a normal distribution with a mean of zero and standard deviation of one. Equation 1 shows this normalization where $\bar{x}$ is the mean and s is the standard deviation of data.

$$\hat{x} = (x - \bar{x})/s \tag{1}$$

This step is important because if the range of one variable is much larger than the range of other variables, it will dominate the clustering and prediction process. By normalizing all variables to the same scale, their contribution in the clustering and prediction models is homogenized.

3 Clustering

Hierarchical clustering technique is chosen for clustering earthquakes because unlike k-means technique it is not sensitive to initial seeds [11]. The distance between two clusters during hierarchical clustering can be calculated using different methods. Average-link method is chosen here because unlike single-link and complete-link methods it is less sensitive to outliers. However, both advantages (not being sensitive to initial seeds and being less sensitive to outliers) come with computational cost [12; 13].

The earthquakes are clustered in 10 classes. If a class contains only one earthquake, that earthquake is eliminated and the clustering process is repeated until each cluster contains more than one earthquake. This iterative elimination process helps filter out outliers.

- Clustering Based on Non-spatial Variables

The earthquakes are clustered based on their magnitude, depth and occurrence date, i.e., earthquakes which have close magnitudes, depths and occurrence dates are more probable to be in the same cluster. At the first iteration, three clusters contained only one earthquake. These three earthquakes were removed and the clustering process was repeated. In the second iteration, there was one cluster with one earthquake. This earthquake was eliminated. In the third iteration, all clusters contained more than one earthquake.

- Clustering Based on Spatial Variables

The earthquakes are clustered based on their location (longitude and latitude). There is no need to normalize the variables (columns) for this clustering because the distances in longitude and latitude are compatible with and compensate each other. The resultant clusters contained more than one earthquake in the first iteration.

- Clustering Based on All Variables

The earthquakes are clustered based on their magnitude, depth, occurrence date, longitude, and latitude. At the first iteration, three clusters contained only one earthquake. These three earthquakes were removed and the clustering process was repeated. In the second iteration, all clusters contained more than one earthquake.

4 Earthquakes Clusters and Magnitude Prediction

Since the prediction model is developed for earthquakes in each cluster independently and separately, earthquakes in each cluster are moved to a new dataset. Thus, the number of sub-datasets is equal to the number of clusters, the union of sub-datasets is the original dataset and the intersection of sub-datasets is empty.

One of the variables required for predicting the magnitude of earthquakes is the number of days passed since the last earthquake. We call this variable “Gap” and add it to each sub-dataset separately. To calculate Gap, first the sub-dataset is ordered in an ascending order based on NDays. NDays is representative of the earthquake occurrence date as the number of days passed since 1900. Gap for an earthquake is equal to its NDays subtracted by the NDays of its immediate predecessor in the sub-dataset. Since Gap cannot be calculated for the first earthquake, it is removed from the sub-dataset.

A prediction model is developed for each sub-dataset to predict the magnitude based on other predictors. The prediction model for the sub-dataset is evaluated using 10-fold cross validation and root mean square error (RMSE) is calculated to evaluate the accuracy of the prediction model. As mentioned before, there are 10 sub-datasets for each dataset, each includes earthquakes of a specific cluster. Consequently, there will be 10 different prediction models with 10 different RMSEs. However, to achieve a single RMSE for the entire dataset (including 10 sub-datasets), the weighted average of these 10 RMSEs is calculated. The weight is the number of earthquakes in the sub-dataset. Three different prediction models are experimented: decision tree, SVM, and kNN.

5 Results

Figs. 2, 3, and 4 show the size of each cluster (logarithmic scale) in three different clusterings explained in Section 3. The cluster sizes are closer to each other in Fig. 3 compared to the other two cases in Figs. 2 and 4. This observation shows that earthquakes are distributed in a few clusters almost uniformly in terms of their locations, though, in terms of their magnitude, depth and occurrence date, most earthquakes (80%) are in one class while the rest of them are distributed in nine other classes.

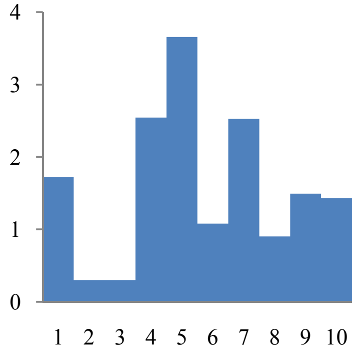

Fig. 2. Size of each cluster (log 10 scale) when clustering based on non-spatial variables.

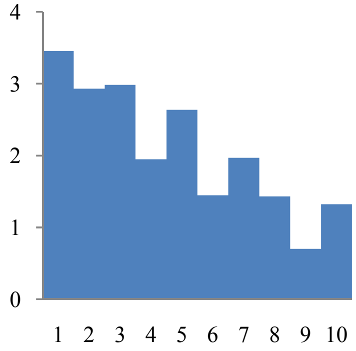

Fig. 3. Size of each cluster (log 10 scale) when clustering based on location.

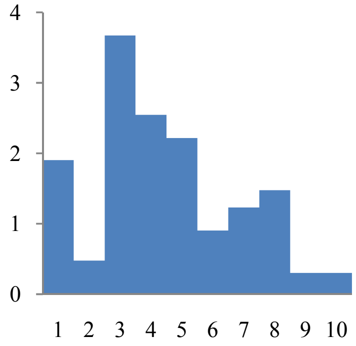

Fig. 4. Size of each cluster (log 10 scale) when clustering based on all variables.

Figs. 5, 6, and 7 show the spatial distribution of earthquakes colored based on their clusters. In Fig. 5, earthquakes are colored based on non-spatial attributes (magnitude, depth, and occurrence date) clusters. In Fig. 6, earthquakes are colored based on location clusters. In Fig. 7, earthquakes are colored based on all variables clusters. Lines in these figures are faults. When earthquakes are clustered based on their non-spatial attributes, the geographical distribution of clusters seems random and does not follow the location of faults. In other words, the earthquakes of one cluster may be located in different parts of the region. On the other hand, when earthquakes are clustered only based on their locations, clusters follow the faults. This is compatible with the concept that earthquakes are stacked around faults [10]. Finally, when both spatial and non-spatial attributes of earthquakes are taken into account for clustering, additional factors affect the geographical distribution of clusters. In areas with low seismicity (lower number of earthquakes), most earthquakes belong to one cluster and there are rarely earthquakes of other clusters. However, in areas with high seismicity, such as southwestern U.S., earthquakes of different clusters are stacked together. This observation shows that in areas with low seismicity, geographical location of earthquakes dominates the clustering but in seismically active areas, with dense historical earthquakes, other non-spatial attributes dominate the clustering.

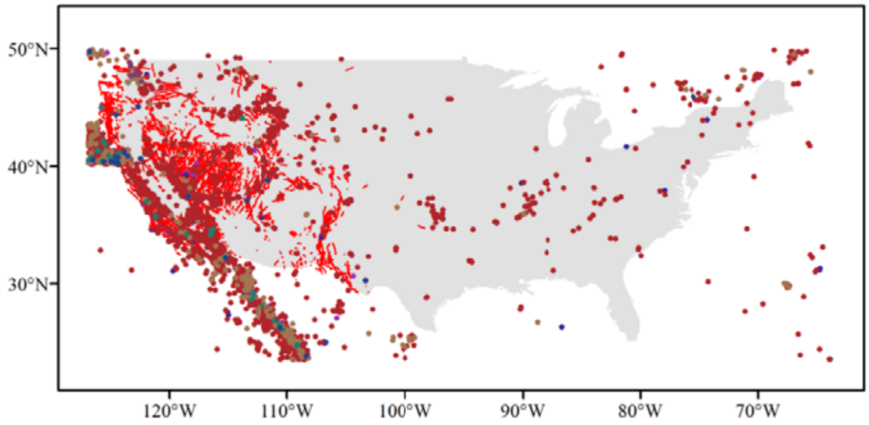

Fig. 5. Clustering earthquakes based on non-spatial attributes.

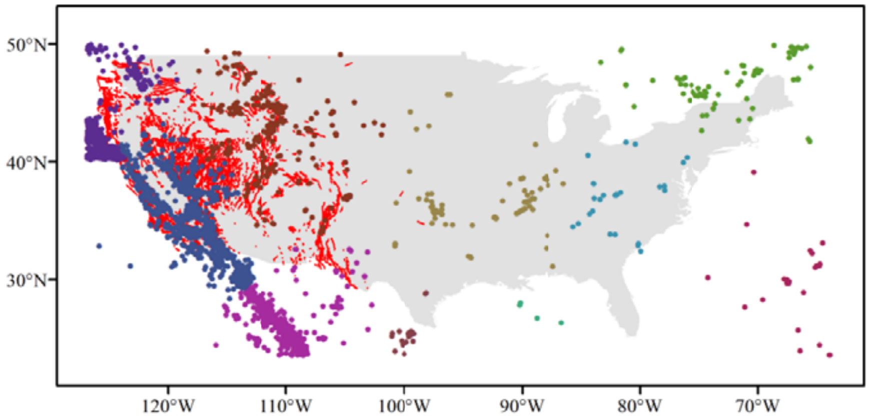

Fig. 6. Clustering earthquakes based on location.

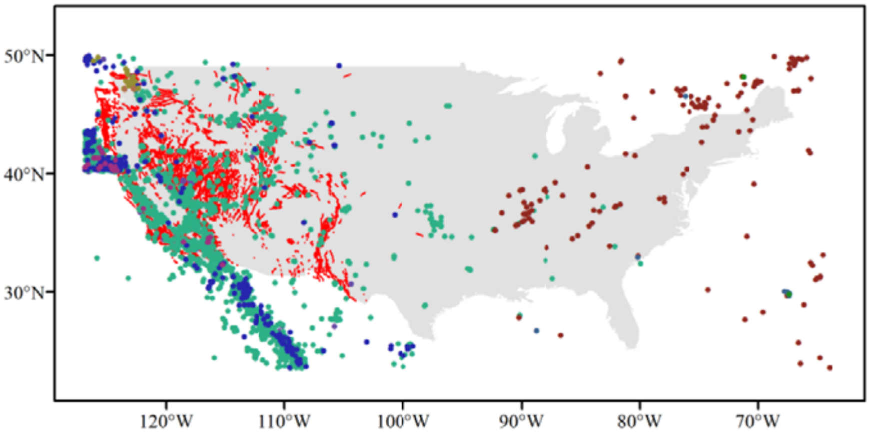

Fig. 7. Clustering earthquakes based on all attributes.

The results for different prediction models and clustering criteria are shown in Table 2. The RMSE (last column in the table) indicates the accuracy of the prediction model. This RMSE is obtained through 10-fold cross validation. Following are the steps to calculate the RMSE of 0.329 in the first row of Table 2:

- Obtain the earthquake clusters (10 in total) based on magnitude, depth, and occurrence date.
  - Develop one prediction model for each of these 10 clusters.
  - Evaluate each prediction model using 10-fold cross validation and calculate a RMSE.
- Calculate the weighted average of ten RMSEs which is 0.329.

The accuracy measure (RMSE) is affected by variables considered for clustering and the prediction model. According to Table 2, kNN is obviously not a good prediction model because its RMSE is much larger than the RMSEs for the other two prediction models. Decision tree has a slightly smaller RMSE than SVM. Besides, decision tree is a much faster prediction model than SVM [11; 12; 13].

When earthquakes are clustered based on their non-spatial attributes (magnitude, depth, and occurrence date) and only depth and Gap are used to predict the magnitude, the least RMSE (highest accuracy) is achieved. When earthquakes are clustered based on their location and only their location is used to predict the magnitude, the worst accuracy is observed. When both non-spatial attributes and location of earthquakes are considered for clustering and depth, Gap and location of earthquakes are used together to predict their magnitude, the observed RMSE is close to the average of the two previous cases. With these results, it can be concluded that taking the locations of earthquakes into account has an adverse effect on accuracy of the prediction model.

Table 2. Results of different prediction models after clustering.

Clustering criteria	Variables used in prediction of magnitude	Prediction model	RMSE
Magnitude, depth, and occurrence date	Depth and Gap	Decision tree	0.329
		SVM	0.339
		kNN (k=10)	6.905
Longitude and latitude	Longitude and latitude	Decision tree	0.541
		SVM	0.564
		kNN (k=10)	8.262
Magnitude, depth, occurrence date, longitude, and latitude	Depth, Gap, longitude and latitude	Decision tree	0.4386
		SVM	0.460
		kNN (k=10)	11.068

Table 3 shows the RMSE of magnitude prediction models without clustering earthquakes. In other words, all earthquakes are considered as one cluster. Comparing the accuracy (RMSE) of prediction models in Tables 2 and 3 shows the effect of clustering on the accuracy of prediction models. According to these two tables, clustering earthquakes decreases the RMSE (improves the accuracy of the magnitude prediction model) by 30% on average over all cases. This observation confirms that clustering earthquakes has been partly successful in generating homogeneous seismic sources. Clustering earthquakes based on their non-spatial attributes (magnitude, depth, and occurrence date) results in the least RMSEs over all different prediction models compared to clustering earthquakes based on spatial or all criteria. In other words, clustering earthquakes based on non-spatial attributes produces the most homogenous hazard zones.

Table 3. Results of different prediction models without clustering.

Variables used in prediction of magnitude	Prediction model	RMSE
Depth and Gap	Decision tree	0.502
	SVM	0.527
	kNN (k=10)	10.143
Longitude and latitude	Decision tree	0.545
	SVM	0.571
	kNN (k=10)	13.884
Depth, Gap, longitude and latitude	Decision tree	0.505
	SVM	0.526
	kNN (k=10)	13.765

6 Conclusions and Future Directions

The most accurate earthquake magnitude prediction model is obtained when the earthquakes are clustered based on their depth, occurrence date and magnitude and the predictors in the prediction model are depth and Gap (number of days passed since the last earthquake in a specific cluster). Adding location of earthquakes to the clustering criteria and predictors weakens the prediction model. Among the three prediction models experimented in this work to predict the magnitude of earthquakes, decision tree was 95% more accurate than kNN and 4% more accurate than SVM in terms of RMSE.

Clustering earthquakes reduced all RMSEs by 30% on average which shows clustering earthquakes, as proposed in this work, is a potential approach in producing homogenous seismic sources which can later be used for producing hazard maps. It is also shown that clustering earthquakes based on their non-spatial attributes (magnitude, depth, and occurrence date) produces the most homogenous seismic sources compared to other clustering criteria.

Clustering earthquakes based on their non-spatial attributes (magnitude, depth, and occurrence date) resulted in one large cluster and many small clusters. Clustering inside the largest cluster was considered, discarding the other small clusters. However, the results are not shown in this article because it resulted in one very large cluster and other very small clusters. This observation implies that the actual clusters are circularly nested inside each other and cannot be separated using regular k-means or hierarchical clustering approaches. However, this hypothesis requires further investigation and is a future research direction.

References

1. Scholz, C. H. (2010). Large earthquake triggering, clustering, and the synchronization of faults. Bulletin of the Seismological Society of America , 100 (3), 901-909.
2. Anderson, J. G., & Nanjo, K. (2013). Distribution of Earthquake Cluster Sizes in the Western United States and in Japan. Bulletin of the Seismological Society of America , 103 (1), 412-423.
3. Cornell, C. (1968). Engineering seismic risk analysis. Bulletin of Seismological Society of America , 58, 1583–1606.
4. Reiter, L. (1990). Earthquake hazard analysis, Issues and insights. New York: Columbia University Press.
5. Anagnos, T., & Kiremidjian, A. S. (1988). A review of earthquake occurrence models for seismic hazard analysis. Probabilistic Engineering Mechanics , 3 (1), 3-11.
6. Hashemi, M., Alesheikh, A. A., & Zolfaghari, M. R. (2013). A spatio-temporal model for probabilistic seismic hazard zonation of Tehran. Computers & Geosciences , 58, 8-18.
7. Erdik, M., Biro, Y. A., Onur, T., Sesetyan, K., & Birgoren, G. (1999). Assessment of earthquake hazard in Turkey and neighboring regions. Annali Di Geofisica , 42 (6), 1125-1138.
8. Erdik, M., Demircioglu, M., Sesetyan, K., Durukal, E., & Siyahi, B. (2004). Earthquake hazard in Marmara Region, Turkey. Soil Dynamics and Earthquake Engineering , 24, 605–631.
9. Zmazek, B., Todorovski, L., Džeroski, S., Vaupotič, J., & Kobal, I. (2003). Application of decision trees to the analysis of soil radon data for earthquake prediction. Applied Radiation and Isotopes , 58 (6), 697-706.
10. Hashemi, M., & Alesheikh, A. (2011). Spatio-temporal analysis of Tehran’s historical earthquakes trends. In Proceedings of Advancing Geoinformation Science for a Changing World (pp. 3-20). Utrecht, Netherlands: Springer.
11. Ledolter, J. (2013). Data mining and business analytics with R. Wiley.
12. Conway, D., & White, J. (2012). Machine learning for hackers. O'Reilly Media.

13. Liu, B. (2007). Web data mining: exploring hyperlinks, contents, and usage data. Springer.

14. United States Geological Survey (USGS). (2015). Retrieved from http://earthquake.usgs.gov/earthquakes/search/

15. United States Geological Survey (USGS). (2015). Retrieved from http://earthquake.usgs.gov/earthquakes/feed/v1.0/glossary.php