Earthquake Cluster Analysis: K-Means Approach

R.K. Kamat[1] and R.S. Kamath[2*]
[1]Department of Electronics, Shivaji University, Kolhapur 416 004
[2]Department of Computer Studies, Chhatrapati Shahu Institute of Business Education and Research, Kolhapur 416004
***Corresponding author: E-Mail: rskamath@siberindia.edu.in**

ABSTRACT

We report unsupervised classification of past ten years seismic events occurred in India based on their spatial location and magnitude using k-means clustering. We useearthquake dataset derived from European-Mediterranean Seismological Center in this work.With the proliferation of efficient analysis, clustering is a time-honored and well understood process for extracting meaningful patterns from large incoherent data sets. We demonstrate the k-means clustering as a tool for analyzing seismic datasetstogether with visualization for interpreting the results. This work exhibits performance of k-means clustering by tuning number of clusters and iterations. Performance is evaluated with reference to within clusters sum of square errors. The present investigation derives six earthquake clusters and thus depicts that k-means has the potential to exhibit the unsurpassed tool for earthquake cluster analysis.

KEY WORDS: k-means, clustering, earthquake, European-Mediterranean Seismological Center.

1. INTRODUCTION

Earthquake clustering is a vital aspect of seismicitywith signatures in space, time, and magnitude domains that provide key information on earthquake dynamics (David, 2016). An early forecast of earthquake is a challenging task since it is a characteristic catastrophe. Most earthquakes related forecasting focuses on minimizing the danger connected with tremors, by evaluating the intermingling of seismic risk and the weakness of a given territory.

Many researchers have carried out the research for the study and analysis of earthquake clusters. Dzwinel (2016), have reported cluster analysis of earthquake events occurred in Japan during the period 1997-2003. Authors have analyzed the cluster patterns in the data space of seismic events using an agglomerative clustering technique (Witold Dzwinel, 2016). Musmeci and Jones have explained a simple space time clustering model with explicit use of space and time intensity for historical earthquakes of Italy (Musmeci, 1992). Zaliapin and Zion have presented statistical analysis of seismicity for the analysis of earthquake clusters (Ilya Zaliapin, 2013). They have used nearest neighbour distances of earthquake events in space, time and energy area. Dataset consists of details of earthquake events occurred in California for the time period 1981 to 2011. The classification detected clusters into several major types. Yet another paper by Vecchio et al have investigated and reported the temporal distribution of earthquakes with common statistical properties (Vecchio, 2008). Molchan & Romashkova have presented earthquake prediction based on empirical analysis of seismic rate using the M8 algorithm (Molchan, 2010). Yet another study by Preethi & Santhi explains time series analysis based on fuzzy optimization for earthquake prediction (Preethi, 2011). Kalita have presented a soft computing approach for the recognition of Earthquake Precursor from low latitude total electron content profiles (Kalita, 2012).

In the backdrop of the research endeavours portrayed above, the present paper demonstrates the knowhow to find spatial patterns of seismic activities. This is accomplished by the cluster analysis which refers to a series of techniques that allow the subdivision of anearthquake dataset into subgroups, based on their similarities. The motivation to earthquake clustering is the fact that, besides reducing the cost of the algorithm, the use of representatives makes the process easier to understand and aids in decision making process. The present paper reports earthquake cluster analysis using k-means clustering. The dataset with 1657 seismic events of India occurred between 1st January, 2005 and 31 December, 2015 is selected for analysis. Each seismic event specifies year, month, day, time, latitude, longitude, magnitude, depth.Since our focus is to derive earthquake clusters based on space and size rather than the time of the earthquake we have filtered out both the date and time from the dataset.

The rest of the paper is structured as follows; after a brief introduction, the second section deals with the theory of k-means clustering. The third section portrays computational details whereas fourth section reports earthquake cluster analysis. The conclusion at the end divulges aptness of the k-means clustering for analysing the seismic tremors.

K-Means clustering: theoretical considerations: The K-means clustering identifies a collection of k clusters using a heuristicsearch starting with a selection of k randomly chosen clusterseach of which in the beginning represents a cluster mean (Kamath, 2016). For each of the remaining data items, one of them is doled out to the most comparable bunch, in light of the separation between the data item and the group mean.Then, it computes the new mean for each cluster.Same process is iterated until the criterion function converges.

Cluster analysis is based on measuring similarity between objects by computing the distance between each pair (Kamath, 2016). The similarity is measured with respect to the mean value of thedata items in a cluster, which can be viewed the cluster's centroid. Typically, the square-error criterion is used for evaluating performance of k-means clustering and is defined as per equation (1).

$$E = \sum_{I=1}^{K} \sum P \in Ci\,|p - mi|^{2} (1)$$

Where E is the sum of the square error for all seismic events in the data set; p is the point in data space representing a given event; and the mean of cluster C_i ism_i. In other words, for each seismic event in each cluster, the distance from the event to its cluster center is squared, and the distances are summed. This principle tries to construct the resulting k clusters as compacted and as separate as possible.

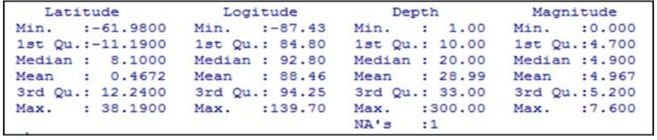
2. METHODS & MATERIALS

Computational details: The present investigation of visualization and analysis of earthquake clusters is simulated in R (Kamath, 2016). R is a language and environment for statistical computing and graphics. The dataset used in the present investigation includes 1657Indian earthquake eventsoccurred in past 10 years. Fig. 1 gives basic statistical computational details of dataset used. This section explains computational details of k-means clustering adopted in the preset research.

We clustered seismic events based on their geographical location and magnitude. A single seismic event E_i can be represented as a multidimensional data vector and is defined as per equation (2).

$$E_i = [X_i, Y_i, D_i, M_i] \qquad (2)$$

Where X_i – Latitude, Y_i- Longitude, D_i– Depth and M_i- Magnitude of seismic event



```
    Latitude             Logitude           Depth             Magnitude
 Min.   :-61.9800   Min.   :-87.43   Min.   :  1.00   Min.   :0.000
 1st Qu.:-11.1900   1st Qu.: 84.80   1st Qu.: 10.00   1st Qu.:4.700
 Median :  8.1000   Median : 92.80   Median : 20.00   Median :4.900
 Mean   :  0.4672   Mean   : 88.46   Mean   : 28.99   Mean   :4.967
 3rd Qu.: 12.2400   3rd Qu.: 94.25   3rd Qu.: 33.00   3rd Qu.:5.200
 Max.   : 38.1900   Max.   :139.70   Max.   :300.00   Max.   :7.600
                                     NA's   :1
```

Figure.1. Basic statistical computational details earthquake dataset

The K-means algorithm takes the inputparameter, k, and partitions a set of n data items into k clusters in such a waythat the resulting intra cluster likeness is high but the inter cluster likenessis low.In the present study the k-means algorithm is applied on normalized data to achieveeffective result. The experiment is tuned by varying number of clusters from 2 to 15 by keeping number of iterations 15, as constant. The entire experiment is summarized in table 1. Performance is evaluated with reference to within clusters sum of square errors and BSS/TSS. Sum of square error explained with equation (1). Sum of Squares (SS), is the usual decomposition of deviance in deviance "Between" and deviance "Within". Ideally a clustering that has the properties of internal cohesion and external separation, i.e. the BSS/TSS ratio should approach 1.

Table.1. Performance evaluation for accuracy of K-means clusteringconfiguration

No. of Clusters	Within cluster sum of squares by cluster	between_SS / total_SS
2	31.15783 15.10191	60.0 %
3	16.389272 11.982872 8.572542	68.0 %
4	4.733471 2.040574 4.897528 16.110757	76.0 %
5	4.649685 6.118403 4.134389 2.040574 4.047918	81.8 %)
6	4.6496855 2.0405741 5.1502686 3.4541328 2.0479033 0.3912264	84.7 %
7	1.9780914 2.1947423 1.1411086 0.7546765 3.8704839 1.5757250 4.5851779	86.1 %
8	2.5049905 1.5746821 2.1947423 0.3912264 2.9518834 0.7711393 1.9346728 1.1411086	88.4 %
9	0.3912264 1.3183183 2.1466504 0.7749200 2.9681837 0.8929310 1.4369598 1.2193797 1.1411086	89.4 %
10	2.0405741 0.5381925 1.2652113 1.1202370 0.3378818 0.7159081 1.1047782 1.0062268 0.9368665 4.2864999	88.4 %
11	0.7970179 1.1411086 2.1466504 0.3106592 1.1191151 1.0040586 1.83787730.7749200 0.8337177 0.2911713 0.6490798	90.6 %
12	0.6721635 0.4460212 4.2538234 0.5694455 1.1238042 0.4820415 1.0253658 0.3052848 2.0405741 0.3276057 0.2911713 0.8186427	89.3 %
13	0.3678312 0.8057317 0.9835533 1.4153276 2.9812865 0.5359296 0.3131835 0.6155286 0.3727819 0.4697059 0.3378818 0.2221251 0.7022379	91.2 %
14	0.5134291 0.5381925 0.7720666 0.7557681 1.1047782 0.3956972 0.5571570 0.5192415 0.4544590 0.6495295 0.9075434 0.3378818 0.6625508 0.7159081	92.3 %)
15	0.7439864 0.8914156 0.4009407 0.4421405 0.3399696 0.1086798 0.9866744 0.5663791 0.5512478 0.2133967 0.2403742 0.7244673 1.2639796 1.1567843 0.3378818	92.2 %

We have applied k-means technique fornumber of clusters from 2 to 15 and plotted the number of clusters against the "within clusters sum of squares", which is the parameters ought to be minimized during the clustering process. Fig.2, shows within-cluster sum of squares for different numbers of clusters. Plot reveals that this quantity decreases up to a point 6, and then remains constant.

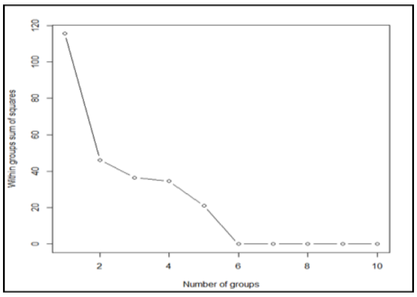
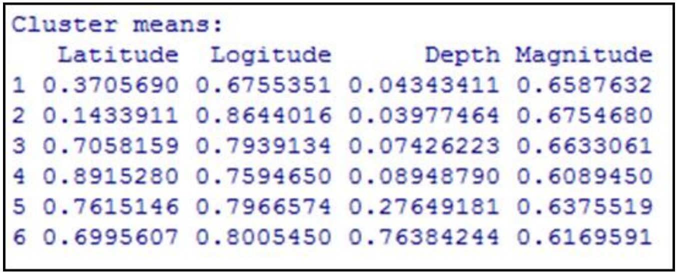


Figure.2. Within-cluster sum of squares for different numbers of clusters

3. RESULTS

Earthquake cluster analysis: With reference to the plot shown in fig.2 and considering within clusters SS, the number of clusters selected for k-means algorithm is 6. The k-means algorithm starts by randomly assigning each seismic event to a cluster, and then it calculates the mean centre of each cluster (Fabio Veronesi, 2016). At this point it calculates the Euclidean distance between each event and the two clusters and reassigns them to a new cluster, based on the closest mean centre, then it recalculates the mean centers and it keeps going until the cluster elements stop changing.

Thus derived k-means clustering results into 6 clusters of sizes 303, 149, 817, 218, 152, 18 and corresponding cluster means are shown in fig. 3. Numerals in the figure are normalized values. The within-cluster variation measures the extent of each event in a cluster, differing from the others in the same cluster. Fig. 4 shows data plot and discriminant plot for 6 clusters derived in the present investigation using k-means clustering. Plot reveals that clusters 2, 3 and 4 are closer where as clusters a, 3, and 6 are farther.



```
Cluster means:
   Latitude  Logitude       Depth Magnitude
1 0.3705690 0.6755351 0.04343411 0.6587632
2 0.1433911 0.8644016 0.03977464 0.6754680
3 0.7058159 0.7939134 0.07426223 0.6633061
4 0.8915280 0.7594650 0.08948790 0.6089450
5 0.7615146 0.7966574 0.27649181 0.6375519
6 0.6995607 0.8005450 0.76384244 0.6169591
```

Figure.3. Cluster means (normalized values) of 6 generated clusters

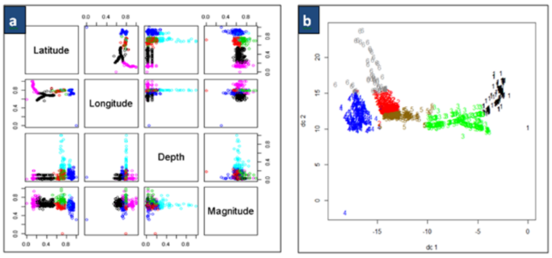


Figure.4. Plots for 6 clusters derived by k-means clustering. (a) Data plot; (b) Discriminant plot

4. CONCLUSION

In the present paper, we have demonstrated earthquake cluster analysis using k-means technique. The dataset with 1657 seismic events of India took place between 1st January, 2005 and 31 December, 2015 is selected for analysis. Present investigation demonstrated cluster analysis by modulating number of clusters. Cluster analysis is accomplished by resorting to a series of techniques that allow the subdivision of a dataset into subgroups, based on their similarities.Thus derived k-means clustering results into 6 clusters of sizes 303, 149, 817, 218, 152, 18. The result strongly suggests that k-means has the potential to exhibit the preeminent tool for earthquake cluster analysis.

5. ACKNOWLEDGEMENT

Dr. R.K. Kamat is gratefully acknowledging the grant-in-aid support received from University Grants Commission, New Delhi under the major research project "Development, testing and prototyping of wireless sensor network for Earthquake early warning system".

REFERENCES

David A, Yuen, Witold Dzwinel, Yehuda Ben-Zion, Ben Kadlec, Visualization of Earthquake Clusters over Multidimensional Space, 2016.

Fabio Veronesi, Cluster analysis on earthquake data from USGS, 2016.

GrahamWilliams, Data Mining with Rattle and R, The Art of Excavating Data for Knowledge Discovery, Springer, New York, 2011.

Ilya Zaliapin and Yehuda Ben-Zion, Earthquake clusters in southern California I, dentification and stability, Journal of Geophysical Research, Solid Earth, 118, 2013, 2847–2864.

Kalita S, Devi M, Barbara A & Talukdar P, Soft Computing Technique for Recognition of Earthquake Precursor from Low Latitude Total Electron Content (TEC) Profiles, International Journal of Computer Applications, 44 (17), 2012, 11-14.

Kamath R.S and Kamat R.K, K-Means Clustering for Analyzing Productivity in Light of R & D Spillover, International Journal of Information Technology, Modeling and Computing (IJITMC), 4 (2), 2016, 55-64.

Kamath R.S, Kamat R.K, Educational Data Mining with R and Rattle, River Publishers Series in Information Science and Technology, River Publishers, Netherland, 2016.

Molchan G, & Romashkova L, Earthquake prediction analysis based on empirical seismic rate, the M8 algorithm. Geophysical Journal International, 183 (3), 2010, 1525-1537.

Musmeci F and Vere-Jones D, A Space-Time Clustering Model for Historical Earthquakes, Ann. Inst. Statist. Math, 44 (1), 1992, 1-11.

Preethi G, & Santhi B, Study on Techniques of Earthquake Prediction. International Journal of Computer Applications, 29 (4), 2011, 55-58.

Vecchio A, Carbone V, Sorriso-Valvo L, De Rose C, Guerra I and Harabaglia P, Statistical properties of earthquakes clustering, Nonlin. Processes Geophys, 15, 2008, 333–338.

Witold Dzwinel, David A.Yuen, Krzysztof Boryczko, Yehuda Ben-Zion, Shoichi Yoshioka, Takeo Ito, Cluster Analysis, Data-Mining, Multi-dimensional Visualization of Earthquakes over Space, Time and Feature Space, 2016.