**Accelerograms selection for time history analyses via clustering by a genetic algorithm compared with K-means method**

Mohsen Shahrouzi[*] and Ehsan Akbari

*Engineering Department, Kharazmi University, Karaj, Iran*

[*]Corresponding author: shahruzi@khu.ac.ir

**Summary**

Linear and nonlinear time-history analyses are necessary tools for many fields of structural/earthquake engineering including vulnerability analyses and response evaluations. In this regard, one of the most important requirements is the selection of proper acceleration time histories as the analysis input. These records can be obtained from natural earthquake records or can be generated synthetically and artificially. The first option is usually preferred to the others since it can provide true information about strong ground shaking characteristics to reflect the source, path, and site effects. These characteristics include several parameters such as earthquake magnitude, epicenter distance, duration, Arias intensity, spectral intensity, soil type and peak ground responses. However, well-known seismic design codes offer their linear design spectra to which the earthquake accelerogram is scaled; instead of direct matching the seismological parameters.

The number of ground motions recorded at a given site is not generally sufficient; thus, it is required to select other earthquakes from real-world catalogues based on their similarity to the site-specific record(s). However, it is a challenging task to deserve such a similarity due to the existence of several criteria affecting the earthquake characteristics. The present work deals with the problem of selecting a number of appropriate ground motions among an available catalogue in order to best match a given site-specific earthquake record.

As an unsupervised solution in this research, clustering techniques were concerned. A set of data in a multi-dimensional space were classified into some clusters in such a way that the objects in every cluster have the most similarity with each other and the least similarity with any object in the other clusters.

Here, both deterministic and non-deterministic approaches in such a clustering problem were employed and compared. The method of *K-means* was selected as a deterministic algorithm due to its popularity and computational efficiency among clustering techniques. A set of input clusters was required to be introduced for this algorithm before its run. However, this method itself was deterministic. Such an initial set is usually chosen by trial-error or randomly. Depending on the initial input clusters, the *K-means* algorithm reveals different local optima instead of global optimum as the clustering result.

In order to overcome such a challenge, a non-deterministic approach was studied as well. In this regard, a *Genetic Algorithm*, GA, was utilized as a well-trusted meta-heuristic among non-deterministic algorithms with the capability of overpassing local results and approaching to the true global optimum. Proper genetic encoding and operators were developed for the current optimal clustering problem. Consequently, a variety of earthquake characteristics were taken into account in the utilized chromosome including magnitude, epicenteral distance, duration, Arias intensity, spectral intensity, soil type, peak ground responses and a specialized error measure. This measure was defined based on how compatible was the response spectrum of the record with the codified design spectrum as the target.

In the present study, average silhouette widths and related profile plot were used to evaluate effectiveness of the proposed clustering processes. Quality of clustering was thus measured taking into account both compactness and separation of the clusters. Consequently, the clustering result corresponding to the maximum amount of the average silhouette bandwidth was announced as the solution. Input-records required for the time-history analyses, were then selected from the cluster which includeed the target site-associated earthquake. Using different numbers of clusters, the proposed genetic search stood well superior to the method of *K-means* in achieving the best clustering for the present problem.