Pipeline Damage Assessment Using Cluster Analysis

S. Toprak[1], E. Nacaroglu[1], O. A. Cetin[1], and A. C. Koc[1]

[1]Department of Civil Engineering, Pamukkale University, Kinikli Campus, Denizli, Turkey; PH +90 2582963352; FAX +90 2582963382; email: stoprak@pau.edu.tr

ABSTRACT

Clustering techniques are used commonly in a variety of engineering and scientific fields although it is rarely used in lifeline earthquake engineering. Cluster analysis deals primarily with the discovery of structures or groupings within data. This paper presents the application of cluster analysis in pipeline damage assessment and identification of high damage areas. Identification of sites where pipeline damage concentrates has special importance because these sites are generally problematic areas and/or pipelines in these sites have some weaknesses. Understanding why damage is high there may contribute and improve future works related to pipeline damage prevention and mitigation. The 1994 Northridge earthquake water distribution pipeline damage in the city of Los Angeles was used herein to illustrate the application of subtractive and fuzzy c-means cluster analysis.

INTRODUCTION

The observations of pipeline damage from past earthquakes close to urban areas indicate that damage may concentrate at particular sites (high damage areas) in the earthquake hit area. This may be result of certain characteristics of the sites and the pipeline system. Existence of fault lines, steep slopes, and weak geotechnical characteristics (e.g., soft clays and liquefiable soils) at sites may cause significant damage to pipelines. Pipeline material, size and joint types are some of the characteristics of a pipeline system which affect the damage potential. Seismic intensity at the pipeline coverage area is another factor which controls the damage level.

Identification of sites where pipeline damage concentrates has special importance because these sites are generally problematic areas and/or pipelines in these sites have some weaknesses. Understanding why damage is high there may contribute and improve future works related to pipeline damage prevention and mitigation. There are different ways of identifying the areas where pipeline damage concentrates. When a human being looks at the map of pipeline damage data (e.g., shown as points at repair locations), he classifies them according to perceived similarities and organizes data into sensible groupings. In this sense, a cluster can be defined as a collection or group of similar objects. Cluster analysis deals with the discovery of structures or groupings within data. Although clustering of data can be achieved manually as described, using clustering techniques have many advantages such as the use of a specified objective criterion consistently to form the groups and

the ability to deal with large number of data sets. The speed, reliability, and consistency of a clustering algorithm in organizing data together constitute an overwhelming reason to use it (Jain and Dubes, 1988).

Water distribution pipeline damage in the city of Los Angeles during the 1994 Northridge earthquake was utilized in this study to illustrate the use of clustering techniques in pipeline damage assessment and identification of high damage areas. O'Rourke and Toprak (1997) presents the largest databases ever assembled in U.S. of spatially distributed transient and permanent ground displacements in conjunction with damage to water supply and distribution lifelines. The 1994 Northridge earthquake caused the most extensive damage to a US water supply system since the 1906 San Francisco earthquake. Three major transmission systems, which provide over three-quarters of the water for the City of Los Angeles, were disrupted. Los Angeles Department of Water and Power (LADWP) and Metropolitan Water District (MWD) trunk lines (nominal pipe diameter ≥ 600 mm) and the LADWP distribution pipeline (nominal pipe diameter < 600 mm) system were damaged. Comprehensive treatment of the earthquake-induced damage to water pipelines and the database developed to characterize this damage can be found at Toprak (1998) and O'Rourke, et al. (1998). In their studies as well as this study, 944 distribution line repairs were identified and used for which there are data pertaining to pipe composition and size. Figure 1 shows the locations of repairs made to the water distribution pipelines of Los Angeles after the 1994 Northridge earthquake. The repair data used herein are

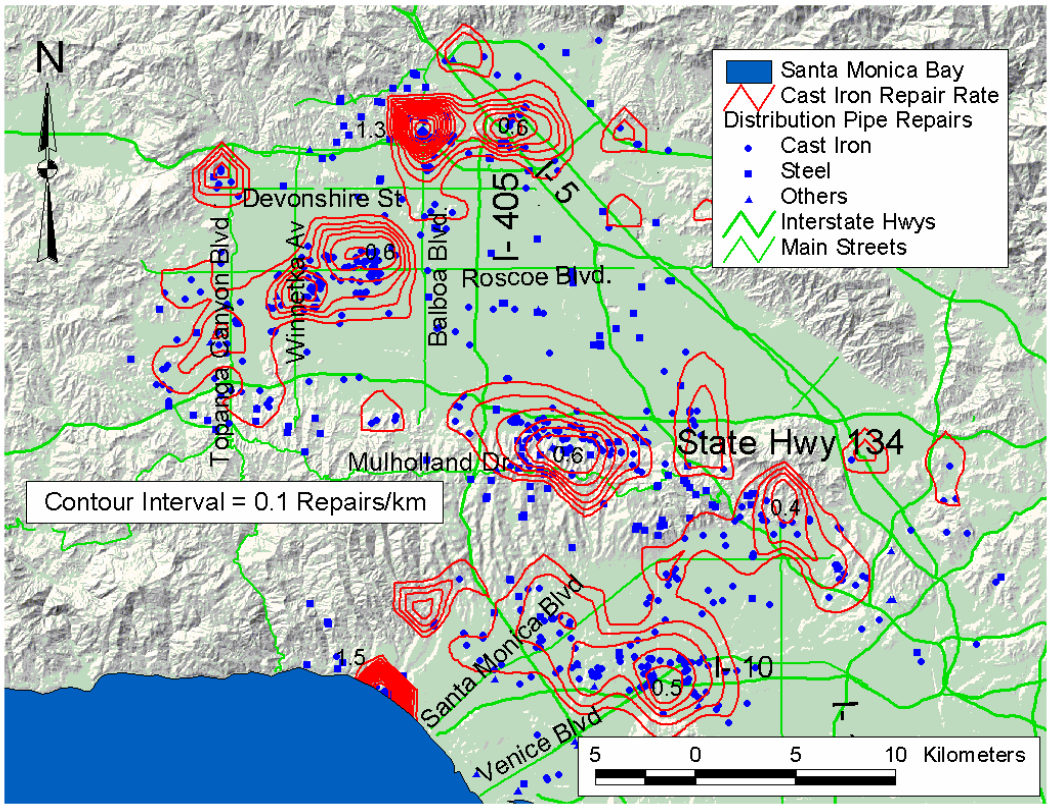

Figure 1. Pipeline repairs and cast iron repair rate contours for the Northridge earthquake (O'Rourke and Toprak, 1997).

slightly different than those used in previous studies as pipe type and pipe size of some repairs have been changed to match the existing pipelines at respective locations as described by Toprak, et al. (2008). The total length of the distribution lines is 10,750 km. About 76%, 11%, 9%, and 4% of the distribution lines are composed of cast iron (CI), steel, asbestos cement (AC) and ductile iron (DI), respectively. Out of 944 distribution line repairs, about 78%, 17%, 3%, 1%, and 1% are cast iron, steel, asbestos cement, ductile iron and other pipe type repairs, respectively.

By looking at Figure 1, one can identify approximately the areas where pipeline damage concentrates. Another method, other than cluster analysis, can be to draw repair rate contours and locate the areas of concentrated contours. Figure 1 presents repair rate contours for CI pipeline damage. The repair rate contours were developed by dividing the map into 2 km x 2 km areas, determining the number of CI pipeline repairs in each area, and dividing the repairs by the distance of CI mains in that area. Contours then were drawn from the spatial distribution of repair rates, each of which was centered on its tributary area. A variety of grids were evaluated, and the 2 km x 2 km grid was found to provide a good representation of damage patterns for the map scale of the figure (Toprak, et al., 1999).

The zones of highest seismic intensity are shown by areas of concentrated contours. In each instance, areas of concentrated contours correspond to zones where the geotechnical conditions are prone either to ground failure or amplification of strong motion (O'Rourke et al., 2001). Each zone of concentrated damage is labeled in Figure 2 according to its principal geotechnical characteristics. In effect, therefore, Figure 1 is a seismic hazard map for the Los Angeles region, calibrated according to pipeline damage during the Northridge earthquake.

CLUSTER ANALYSIS METHODS

Although there are many cluster analysis techniques available, because of space limitations, only two of them will be presented herein: subtractive clustering and fuzzy c-means. The subtractive clustering method assumes each data point is a potential cluster center and calculates a measure of the likelihood that each data point would define the cluster center, based on the density of surrounding data points. A data point with more neighboring data will have a higher opportunity to become a cluster center than points with fewer neighboring data. The algorithm: i) Selects the data point with the highest potential to be the first cluster center ii) Removes all data points in the vicinity of the first cluster center (as determined by radii), in order to determine the next data cluster and its center location iii) Iterates on this process until all of the data is within radii of a cluster center.

Based on the density of surrounding data points, the potential value for each data point is calculated by Chiu, (1994) as follows:

$$P_i = \sum_{j=1}^{n} e^{\left\| -4 x_i - x_j^2 \right\| / R_a^2}$$

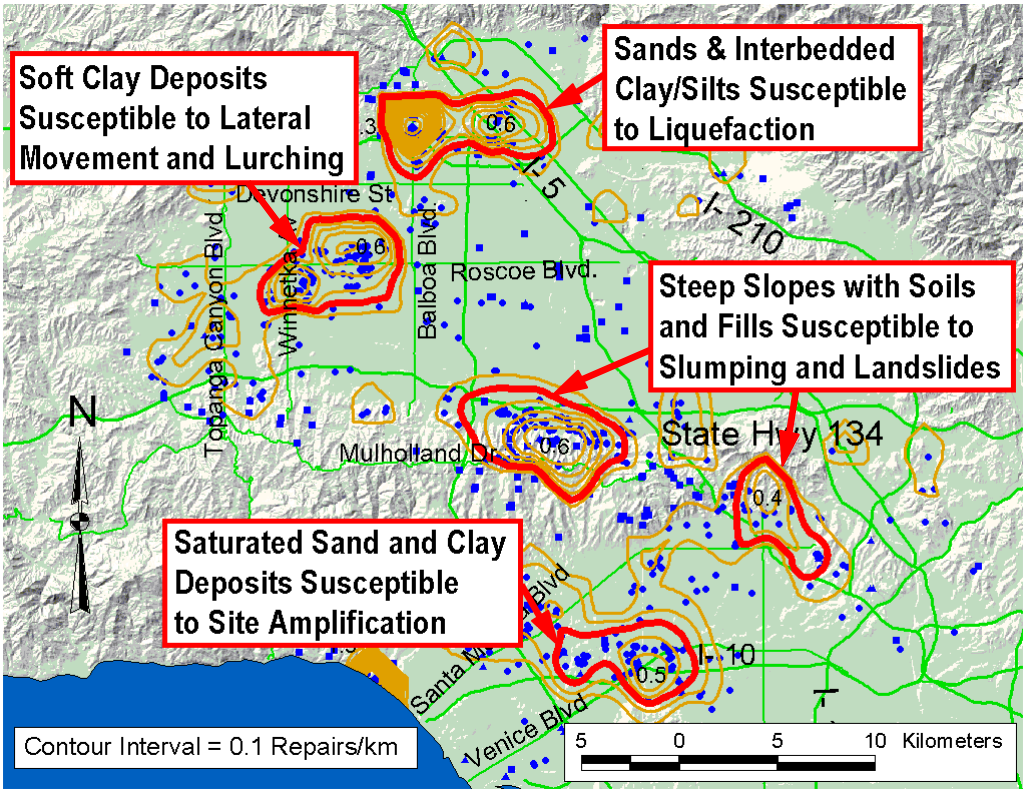

Figure 2. Geotechnical characteristics of the areas of concentrated pipeline damage after the Northridge earthquake (O'Rourke, et al. 2001)

.

where x_i, x_j are data points and R_a is a positive constant defining a neighborhood. Data outside this range have little influence on the potential. After the potential of every data point has been computed, the data point with the highest potential is chosen as the first cluster center. If x_1^* be the location of the first cluster center and P_1^* is its potential value, then the potential of the remaining data points x_i is revised by

$$P_i \Rightarrow P_i - P_1^* e^{\left\|-4\,x_i - x_1^*\right\|^2 / R_b^2}$$

where R_b is a positive constant ($R_b > R_a$). Generally, after the k^{th} cluster center has been obtained, the potential of each data point is revised by

$$P_i \Rightarrow P_i - P_k^* e^{\left\|-4x_i - x_k^*\right\|^2 / R_b^2}$$

Thus, the data points near the first cluster center will have greatly reduced potential, and therefore are unlikely to be selected as the next cluster center. The constant R_b is the radius defining the neighborhood that will have measurable reductions in potential. To avoid obtaining closely spaced cluster centers, R_b is set to

be greater than R_a. Since the parameters R_a and R_b are closely related to each other and R_b is always greater than R_a, the parameter R_b can be replaced by another parameter called the Squash Factor (SF) which is the ratio between R_a and R_b:

$$SF = \frac{R_b}{R_a}$$

The process described above continues until no further cluster center is found. As for whether a data point is chosen as a cluster center or not, there are two parameters involved, the Accept Ratio (AR) and the Reject Ratio (RR). These two parameters, together with the influence range and squash factor, set the four criteria for the selection of cluster centers.

Fuzzy c-means (FCM) is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade. This technique was originally introduced by Bezdek, (1981) as an improvement on earlier clustering methods. It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters. FCM is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} {u_{ij}}^m \left\| x_i - c_j \right\|^2$$

Where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j; x_i is the i^{th} of d-dimensional measured data ; c_j is the d-dimension center of the cluster and $\|*\|$ is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the c_j cluster centers by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}}$$

$$c = \frac{\sum_{i=1}^{N} {u_{ij}}^m . x_i}{\sum_{i=1}^{N} {u_{ij}}^m}$$

This iteration will stop when

$$\max_{ij} \left\{ \left| {u_{ij}}^{(k+1)} - {u_{ij}}^{(k)} \right| \right\} < \varepsilon$$

Where ε is a termination criterion between 0 and 1 and k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

CLUSTER ANALYSIS OF PIPELINE DAMAGE

MATLAB program and existing subroutines with some modifications were used in the analyses (Palm, 2004; Balasko, et al., 2005). Figure 3 shows cluster centers from subtractive cluster analysis for CI pipeline damage of Los Angeles superimposed in CI pipeline repair data. To permit comparison with Figure 2, the four parameters of the method were arranged to get five clusters. Cluster centers were consistent with the damage concentration areas shown in Figure 2.

Figure 4 shows five clusters, which were obtained by fuzzy c-means clustering, superimposed on CI pipeline repairs. Cluster centers are also shown on the figure. Pipeline repairs belonging to the same cluster are shown with the same color. The contours show the membership grades for each cluster. As mentioned previously, each data point belongs to a cluster to some degree that is specified by a membership grade. The contour interval is 0.1. Cluster centers and clusters themselves were in general agreement with the damage concentration areas shown in Figure 2.

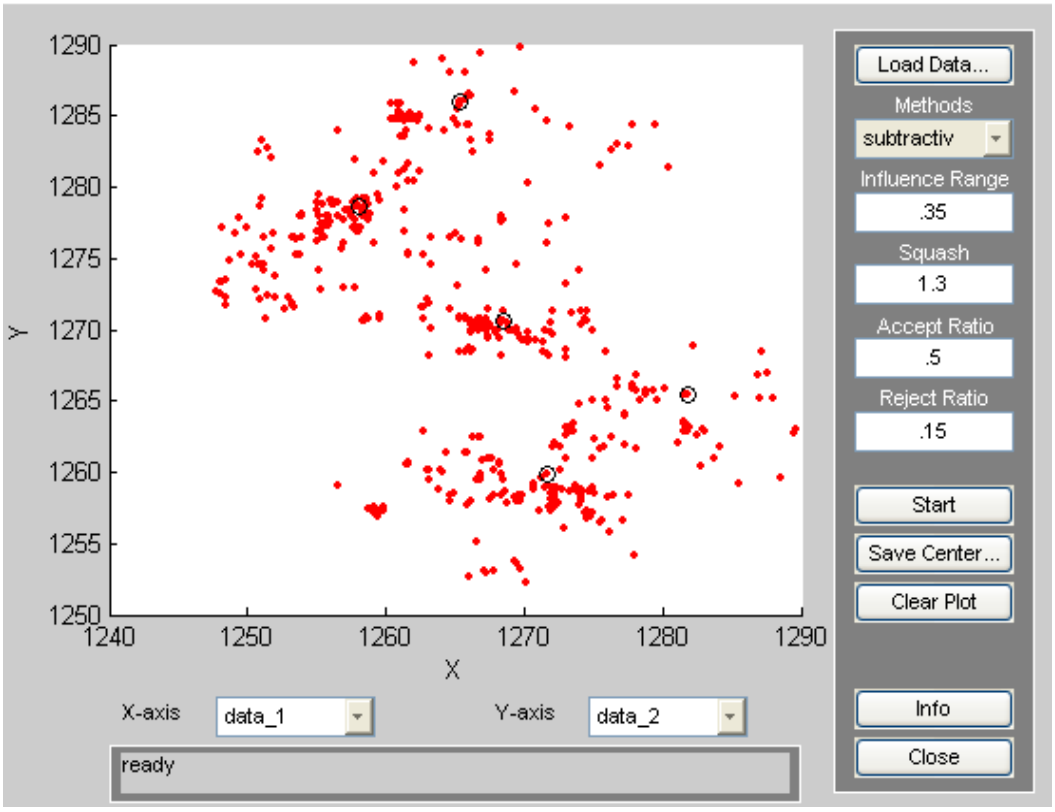

Figure 3. Cluster centers for the pipeline damage using subtractive clustering.

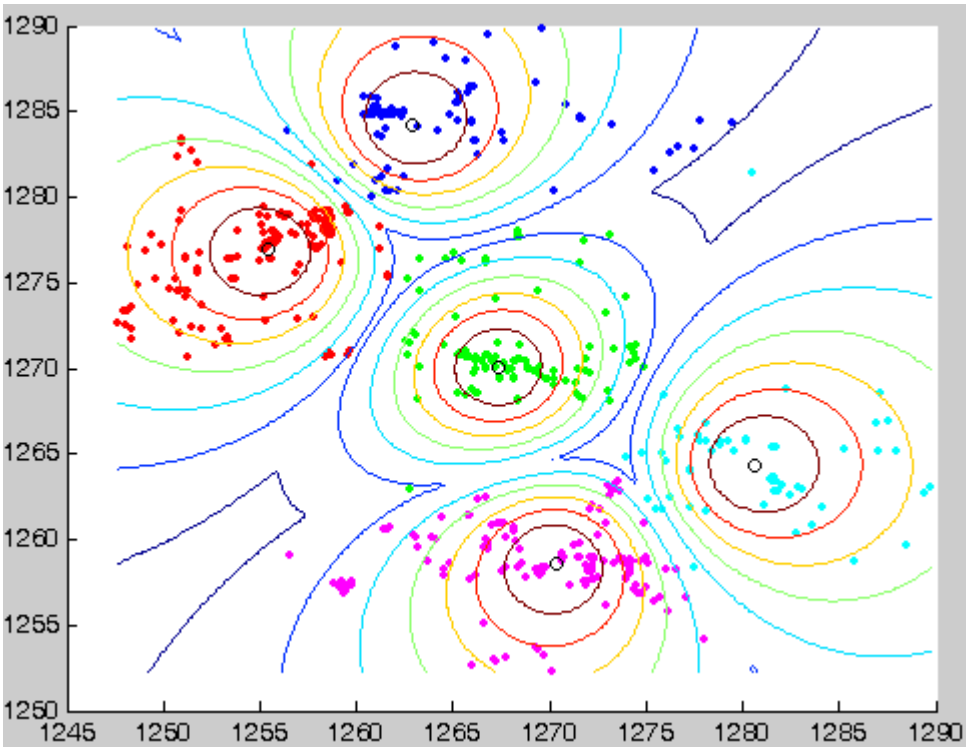

Figure 4. Cluster centers for the pipeline damage using fuzzy c-means cluster analysis.

SUMMARY AND CONCLUSION

This paper presents the application of cluster analysis in pipeline damage assessment and identification of high damage areas. Identification of sites where pipeline damage concentrates has special importance because these sites are generally problematic areas and/or pipelines in these sites have some weaknesses. Understanding why damage is high there may contribute and improve future works related to pipeline damage prevention and mitigation. The 1994 Northridge earthquake water distribution pipeline damage in the city of Los Angeles was used herein to illustrate the application of subtractive and fuzzy c-means cluster analysis. The cluster analysis results presented herein were consistent with previous results obtained from pipeline repair rate contours. A future work will present the effects of different clustering parameters and number of clusters in pipeline damage assessment and identification of high damage areas.

ACKNOWLEDGMENTS

The research reported in this paper was supported by Scientific and Technological Research Council of Turkey (TUBITAK) under Project No. 106M252. Partial grant provided by PAU BAP to attend the conference is acknowledged.

REFERENCES

Balasko, B., Abonyi J., and Feil, B. (2005). *Fuzzy clustering and data analysis toolbox for use with matlab,* Department of Process Engineering, University of Veszprem.

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function*, Plenum Press, New York.

Chiu, S. L., (1994), "Fuzzy model identification based on cluster estimation*", Journal of Intelligent and Fuzzy Systems*, 2, John Wiley & Sons, 267-278.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*, Prentice Hall, Englewood Cliffs, NJ.

O'Rourke, T. D., Stewart, H. E., and Jeon, S-S., (2001). "Geotechnical aspects of lifeline engineering", Proceeding of Institution of Civil Engineers Geotechnical Engineering, 149, Issue 1, Jan. 2001, 13-26.

O'Rourke, T. D. and Toprak, S. (1997). "GIS assessment of water supply damage from the Northridge earthquake." Frost, JD, Editor, Geotechnical Special Publication No. 67, New York, NY: ASCE, 117-131.

O'Rourke, T. D., Toprak S., Sano Y. (1998). "Factors affecting water supply damage caused by the Northridge earthquake.", Proceedings of the 6th US National Conference on Earthquake Engineering, Seattle, WA, USA, 1-12.

Palm, W. J., (1994). *Introduction to Matlab 7 for engineers,* McGraw-Hill, New York.

Toprak, S. (1998). *Earthquake effects on buried lifeline systems,* Ph.D. Thesis, Ithaca, NY, Cornell University.

Toprak, S., Koc, A. C., Cetin, O. A., and Nacaroglu, E. (2008). "Assessment of buried pipeline response to earthquake loading by using GIS.", The 14th World Conference on Earthquake Engineering, Paper 06-0077, October 12-17, 2008, Beijing, China.

Toprak, S., O'Rourke, T. D. and Tutuncu, I. (1999). "GIS characterization of spatially distributed lifeline damage, optimizing post-earthquake lifeline system reliability.", Proceedings, Fifth U.S. Conference on Lifeline Earthquake Engineering, Elliott, W. M. and McDonough, P., Eds., Seattle, WA, August, ASCE, 110-119.