# Use of Subtractive Clustering Analysis in Pipeline Damage Assessment

**S. Toprak, E. Nacaroglu, A. C. Koc, and O. A. Cetin**
*Department of Civil Engineering, Pamukkale University, Kinikli Campus, Denizli, Turkey; email: stoprak@pau.edu.tr*


**ABSTRACT:**
This paper presents the application of subtractive clustering analysis in pipeline damage assessment and identification of high damage areas. Water distribution pipeline damage in the city of Los Angeles during the 1994 Northridge earthquake was utilized to illustrate the use of subtractive clustering. 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. MATLAB program and existing subroutines with some modifications were used in the analysis. Effects of different clustering parameters on cluster centers and number of clusters were determined by using this database.

*Keywords: Clustering, Damage, Pipelines, Seismic*

## 1. INTRODUCTION

Clustering techniques are used commonly in a variety of engineering and scientific fields although it is rarely used in lifeline earthquake engineering. To the best knowledge of authors, Toprak et al. (2009) introduced first time the application of clustering analysis to water pipeline system damage. Cluster analysis deals with the discovery of structures or groupings within data. Although clustering of data can be achieved manually, 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).

Subtractive clustering (Chiu, 1994), is a fast, one-pass algorithm for estimating the number of clusters and the cluster centers in a set of data. The method, as a modified mountain clustering (Yager and Filev, 1994), uses all the data points to replace all the grid points as potential cluster centers. This effectively reduces the number of grid points to number of data points (Chiu, 1994). Effects of different clustering parameters on cluster centers and number of clusters were studied herein by using the database of the City of Los Angeles water supply damage caused by the 1994 Northridge earthquake. Cluster numbers identify the number of sites where pipeline damage is high and cluster centers point to the approximate center of those sites.

## 2. SUBTRACTIVE CLUSTERING

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:

$$P_i = \sum_{j=1}^{n} e^{-4\left\|x_i - x_j^*\right\|^2 / R_a{}^2} \tag{2.1}$$

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

$$P_i \Rightarrow P_i - P_1{}^* e^{-4\left\|x_i - x_1^*\right\|^2 / R_b{}^2} \tag{2.2}$$

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

$$P_i \Rightarrow P_i - P_k{}^* e^{-4\left\|x_i - x_k^*\right\|^2 / R_b{}^2} \tag{2.3}$$

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 $R_b$:

$$SF = \frac{R_b}{R_a} \tag{2.4}$$

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. Accept ratio sets the potential, as a fraction of the potential of the first cluster center, above which another data point will be accepted as a cluster center whereas reject ratio sets the potential, as a fraction of the potential of the first cluster center, below which a data point will be rejected as a cluster center (Mathworks, 2010).

## 3. NORTHRIDGE EARTHQUAKE AND CITY OF LOS ANGELES WATER SUPPLY DAMAGE

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 subtractive clustering technique 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 Toprak, et al. (2008), 944 distribution line repairs were identified and used for which there are data pertaining to pipe composition and size. The repair data used in Toprak, et al. (2009) are 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.

## 4. SUBTRACTIVE CLUSTERING ANALYSIS OF PIPELINE DAMAGE

Figure 1 shows the locations of CI repairs made to the water distribution pipelines of Los Angeles after the 1994 Northridge earthquake. There are 734 CI repairs on the map. MATLAB program and existing subroutines with some modifications were used in the cluster analysis (Mathworks, 2010). Figure 1 shows cluster centers from subtractive clustering analysis for CI pipeline damage of Los Angeles superimposed on CI pipeline repair data. The solid (red) and open circles (black) show the pipeline repair locations and cluster centers, respectively. The numbers used for the four parameters of the method resulted in five clusters. Cluster centers were consistent with the general damage concentration areas as discussed in Toprak, et al. (2009).

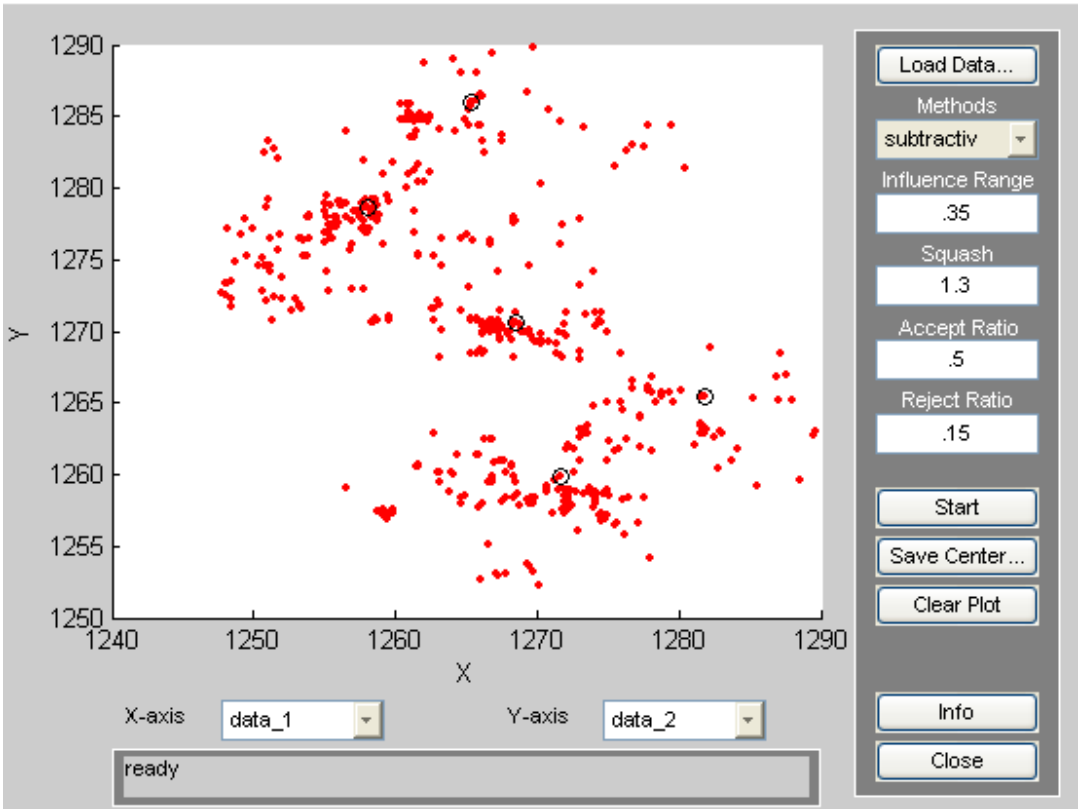



**Figure 1.** Cluster centers for the pipeline damage using subtractive clustering (Toprak, et al., 2009)

Use of different values for influence range (IR), squash factor (SF), the accept ratio (AR), and the reject ratio (RR) results in different cluster centers and cluster center locations. Table 1 shows the cluster numbers for SF=1.25, IR=0.1, and varying AR and RR values. In subtractive clustering, RR values should be smaller than AR values. Although results can be obtained in tabular form as shown in Table 1, it can be much better and compact to show them in graphical form like the one in Figure 2. The plots in the figure obtained for SF = 1.25 and various IR and AR values changing by 0.1 by using the tabular information like the one in Table 1. Figure 2 shows the information contained in 8 tables. The lines in the figure are drawn by using the cluster numbers on the diagonals from the tables. They correspond to any AR value and a RR value 0.1 smaller than that particular AR. For example, the cluster number for AR = 0.5 and RR = 0.4 can be determined from Table 1 as 6. The same value can be obtained from Figure 2 by reading the value corresponding to IR= 0.1 and AR = 0.5. However, the figure can be used to determine any intermediate value by using a property discovered from the tables. It can be noticed that for any column, the cluster numbers are the same below the diagonal value. This property can be used to determine the cluster numbers by using the lines in Figure 2 freely. For example, if one wants to find the cluster numbers for SF=1.25, IR= 0.1, AR= 0.5 and RR=0.2 (which is actually 14), he can read the cluster number corresponding to AR=0.3 (0.1 higher than RR) from the IR = 0.1 line in Fig. 2.

Tablo 1. Cluster numbers determined by subtractive clustering for IR=0.1 ve SF= 1.25

| | RR | 0,1 | 0,2 | 0,3 | 0,4 | 0,5 | 0,6 | 0,7 | 0,8 | 0,9 |
|---|---|---|---|---|---|---|---|---|---|---|
| AR | 0,2 | 22 | | | | | | | | |
| | 0,3 | 22 | 14 | | | | | | | |
| | 0,4 | 22 | 14 | 9 | | | | | | |
| | 0,5 | 22 | 14 | 9 | 6 | | | | | |
| | 0,6 | 22 | 14 | 9 | 6 | 4 | | | | |
| | 0,7 | 22 | 14 | 9 | 6 | 4 | 4 | | | |
| | 0,8 | 22 | 14 | 9 | 6 | 4 | 4 | 4 | | |
| | 0,9 | 22 | 14 | 9 | 6 | 4 | 4 | 4 | 3 | |
| | 1 | 22 | 14 | 9 | 6 | 4 | 4 | 4 | 3 | 1 |

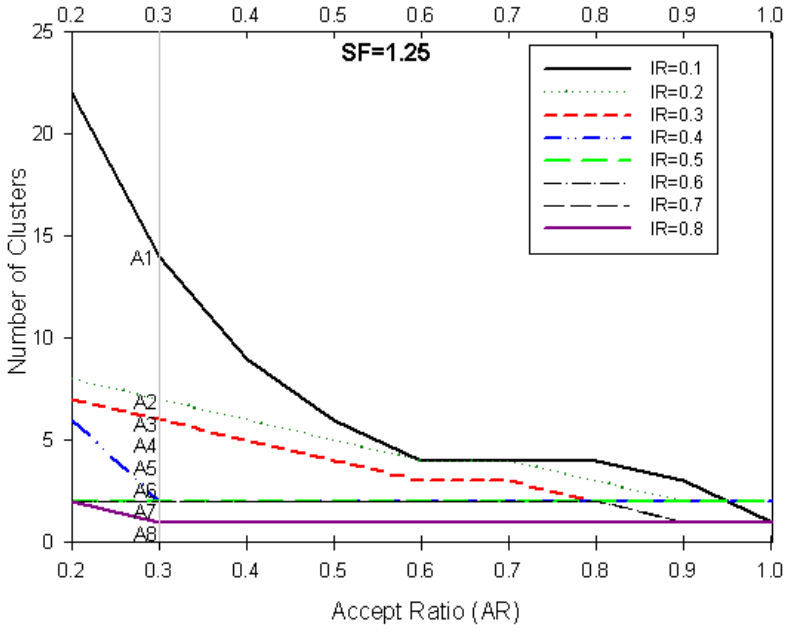
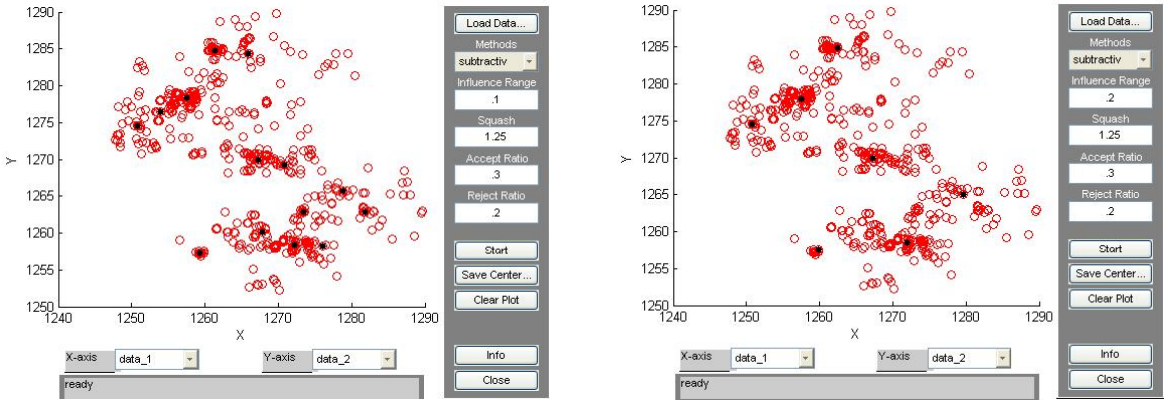


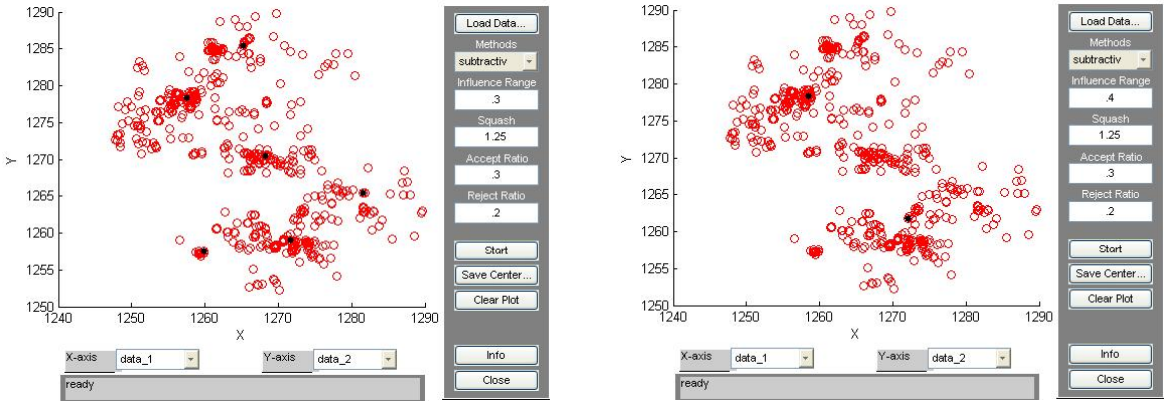

**Figure 2.** Number of clusters for SF=1.25 and changing IR, AR, and RR values




A1- IR=0.1



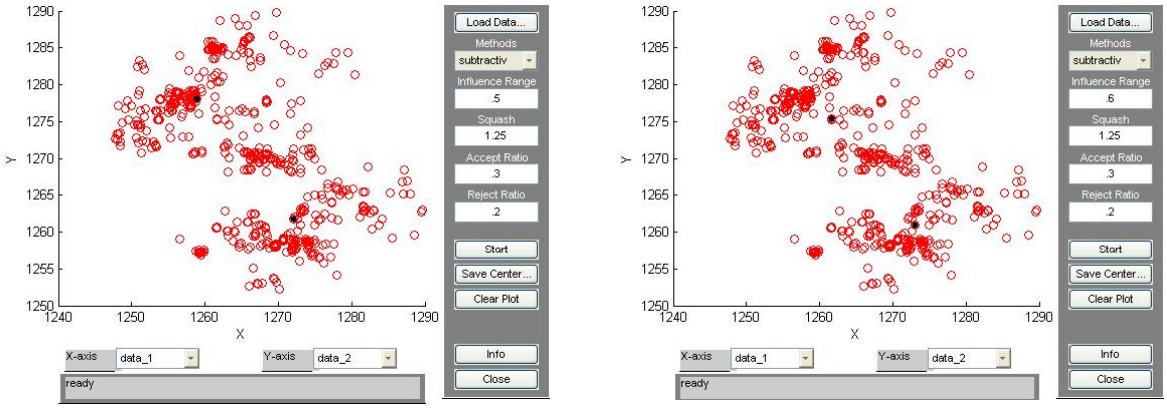

A2- IR=0.2




A3- IR=0.3



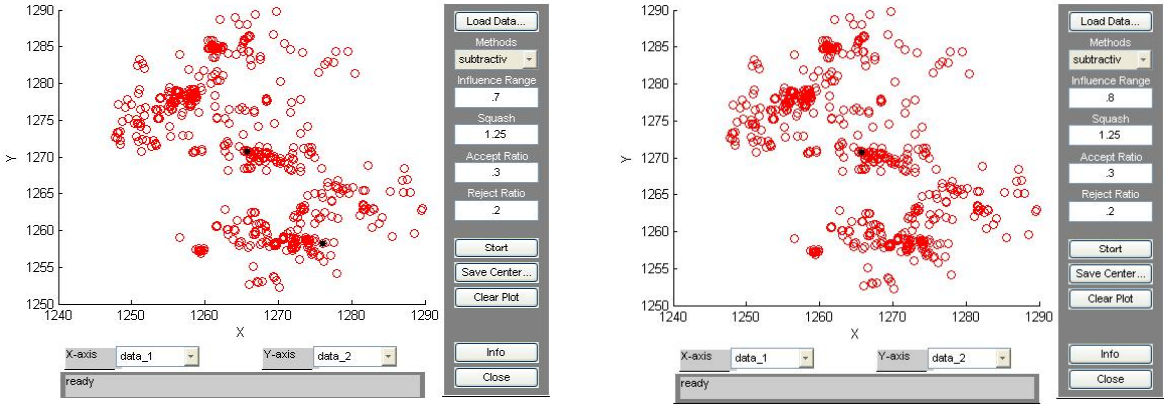

A4- IR=0.4




A5- IR=0.5




A6- IR=0.6



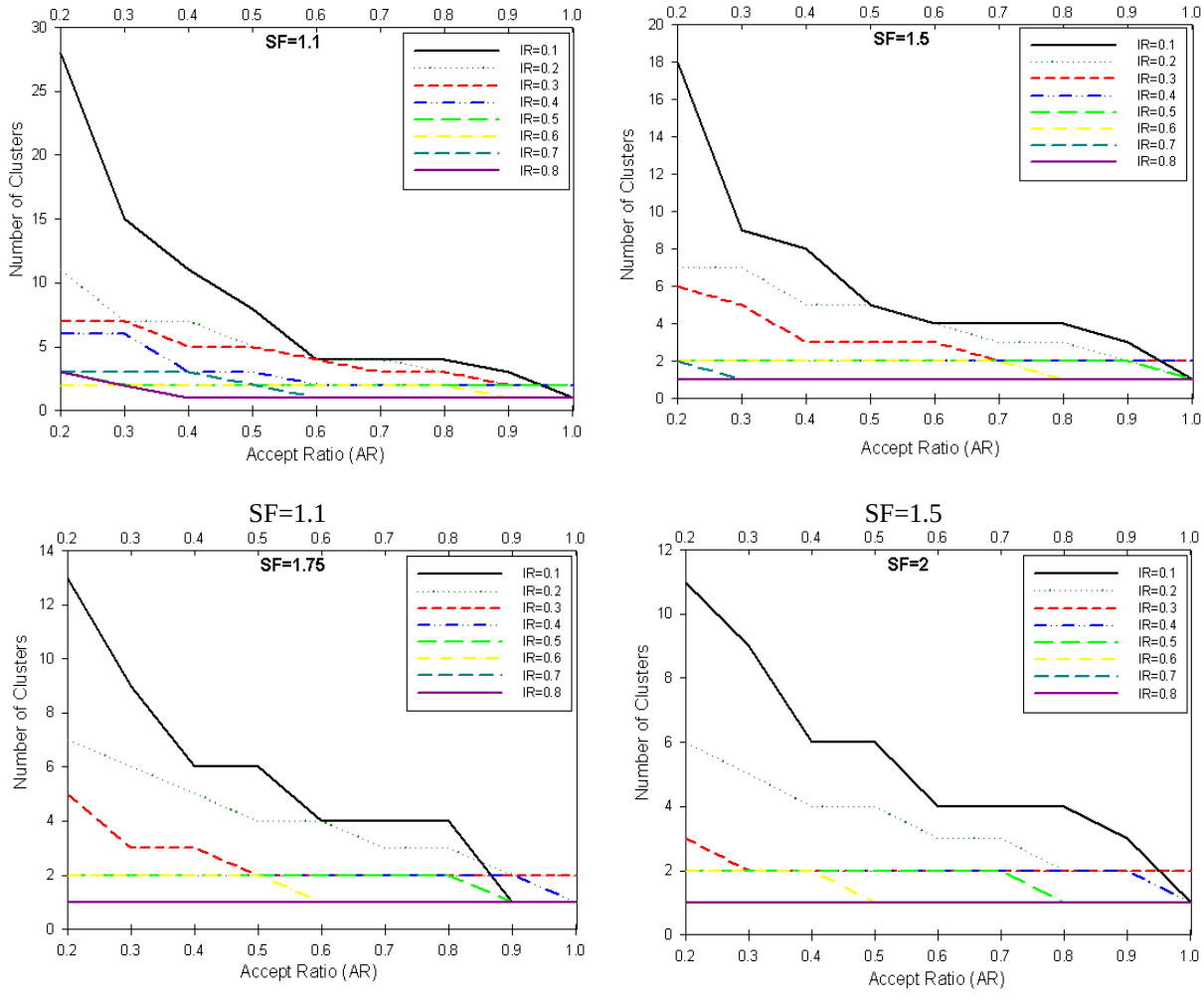

A7- IR=0.7




A8- IR=0.8

**Figure 3.** Cluster centers for SF=1.25 and AR=0.3 in Fig. 2 (A series)

Figure 3 shows the cluster centers superimposed on pipeline damage for SF = 1.25, AR = 0.3, RR= 0.2 and various IR values between 0.1 and 0.8. These cluster centers correspond to A1 to A8 shown in Figure 2. The open (red) and solid (black) circles show the pipeline repair locations and cluster centers, respectively. As IR changes from 0.1 to 0.8, cluster numbers decreased from 14 to 1. Figure 4 presents the results in a similar fashion to Figure 3 and shows the graphs for SF = 1.1, 1.5, 1.75, and 2. In general, the number of clusters is decreased as SF increases for the same values of other three parameters. Considering that at least five clusters are visible in the pipeline damage database, the results illustrate that IR values should be smaller than 0.5 to obtain this many clusters. If SF values get higher (e.g., close to 2), then the IR values should be reduced more (e.g., below 0.3).




SF=1.1 SF=1.5

SF=1.75 SF=2.0

**Figure 4.** Number of cluster centers for the ranges of clustering parameters

## 5. SUMMARY AND CONCLUSIONS

There exist several clustering approaches for data analysis. This paper presents the application of subtractive clustering 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 problematic areas and understanding why damage is high there may contribute future damage prevention and mitigation works. The 1994 Northridge earthquake water distribution pipeline damage in the city of Los Angeles was used herein to illustrate the application of subtractive clustering

analysis. Cluster numbers identify the number of sites where pipeline damage is high and cluster centers point to the approximate center of those sites. Effects of different clustering parameters on cluster centers and number of clusters were determined by using this database. In general, the number of clusters is decreased as SF increases for the same values of other three parameters. Considering that at least five clusters are visible in the existing pipeline damage database, the results illustrate that IR values should be smaller than 0.5 to obtain at least this many clusters. If SF values get higher (e.g., close to 2), then the IR values should be reduced more (e.g., below 0.3). Correlations between subtractive clustering analysis and one other common clustering method, fuzzy c-means cluster analysis will be presented in a future work.

**ACKNOWLEDGEMENT**

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

**REFERENCES**

Chiu, S. L., (1994). Fuzzy Model Identification Based on Cluster Estimation. *Journal of Intelligent and Fuzzy Systems* **2**, 267-278.

Jain, A. K. and Dubes, R. C. (1988). Algorithms for Clustering Data, Prentice Hall, Englewood Cliffs, NJ.

Mathworks (2010). Fuzzy Logic Toolbox™ 2 User's Guide. www.mathworks.com.

O'Rourke, T. D. and Toprak, S. (1997). GIS Assessment of Water Supply Damage from The Northridge Earthquake. *Geotechnical Special Publication* **67**, 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*, 1-12.

R.R.Yager and D.P.Filev. (1994). Generation of Fuzzy Rules by Mountain Clustering. J*ournal of Intelligent and Fuzzy System* **2**, 209-219.

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.

Toprak, S., Nacaroglu, E., Cetin, O. A. and Koc, A. C. (2009). Pipeline Damage Assessment Using Cluster Analysis. *TCLEE 2009: Lifeline Earthquake Engineering in a Multihazard Environment Proceedings of the 2009 ASCE Technical Council on Lifeline Earthquake Engineering Conference* ASCE Conf. Proc. 357, 78.