Journal of Earth Science, *Vol. 25*, *No. 1*, *p. 140–145*, *February 2014* *ISSN 1674-487X*
Printed in China
DOI: 10.1007/s12583-014-0406-x

Clustering Seismic Activities Using Linear and Nonlinear Discriminant Analysis

H Serdar Kuyuk*[1], Eray Yildirim[2], Emrah Dogan[1], Gunduz Horasan[2]
1. *Department of Civil Engineering*, *Faculty of Engineering*, *Sakarya University*, *Sakarya* 54187, *Turkey*
2. *Department of Geophysical Engineering*, *Faculty of Engineering*, *Sakarya University*, *Sakarya* 54187, *Turkey*

ABSTRACT: Identification and classification of different seismo-tectonic events with similar characteristics in a region of interest is one of the most important subjects in seismic hazard studies. In this study, linear and nonlinear discriminant analyses have been applied to classify seismic events in the vicinity of Istanbul. The vertical components of the digital velocity seismograms are used for seismic events with magnitude (M_d) between 1.8 and 3.0 that occurred between 2001 and 2004. Two, time dependent parameters, complexity and S/P peak amplitude ratio are selected as predictands. Linear, quadratic, diaglinear and diagquadratic discriminant functions are investigated. Accuracy of methods with an additional adjusted quadratic models are 96.6%, 96.6%, 95.5%, 96.6%, and 97.6%, respectively with a various misclassified rate for each class. The performances of models are justified with cross validation and resubstitution error. Although all models remarkably well performed, adjusted quadratic function achieved the best success rate with just 4 misclassified events out of 179, even better compared to complex methods such as, self organizing method, *k*-means, Gaussion mixture models that applied to same dataset in literature.
KEY WORDS: discriminant analysis, clustering analysis, self organizing map, *k*-means, Gaussion mixture models.

1 INTRODUCTION

Different methods using various input parameters to categorize earthquakes (EQs) and man-made explosions such as quarry blasts (QBs) have been derived worldwide, each one generally suitable to a particular region. Statistical methods, mainly linear discriminant analysis (LDA) have commonly used in seismic studies (Horasan et al., 2009; Koch and Fah, 2002; Rodgers and Walter, 2002). Artificial intelligence techniques such as artificial neural networks, self-organizing map, adaptive neuro-fuzzy inference system etc. have been employed recently (Kuyuk et al., 2011, 2010; Yildirim et al., 2011; Del Pezzo et al., 2003; Jenkins and Sereno, 2001; Ursino et al., 2001; Musil and Plešinger, 1996). As discriminants, power ratios, amplitude peak ratios, and spectral amplitude ratios etc., which are obtained from time- and frequency-domain analysis of seismograms were utilized in the literature (Gitterman et al., 1998; Wüster, 1993; Bennett and Murphy, 1986).

There is high seismic activity in Marmara region, Turkey where the metropolitan mega-city Istanbul located in the northwest (Fig. 1). Kandilli Observatory and Earthquake Research Institute (KOERI) records and processes the seismic activities for seismic hazard assessment constantly. However, these tasks need to be auto-operated and systematic due to high seismic activity. Operations should be resistant to personnel changes.

Horasan et al. (2009) studied linear discriminant analysis in Marmara region using three parameters: amplitude peak, power and spectral amplitude ratios. They offered an additional variable: origin time of events, because quarry blasts are happening in daytime. Then, Yildirim et al. (2011) demonstrated the use of feedforward neural networks (FFNNs), adaptive neural fuzzy inference systems (ANFISs), and probabilistic neural networks (PNNs) to discriminate between earthquakes and quarry blasts for the region. The input vectors consist of the peak amplitude ratio (S/P ratio) and the complexity value. The success of the developed models on regional test data varies between 97.67% and 100%. Kuyuk et al. (2011) extended discriminants by using four parameters (complexity, spectral ratio, S/P wave amplitude peak ratio and origin time of events) and applied an unsupervised learning approach self-organizing map (SOM); however, they have showed that complexity and S/P parameters correlates better than others. Kuyuk et al. (2012) attempted to approach the same problem with applying *k*-means method and Gaussian mixture model and obtained 95.0% and 96.1% accuracy which is lower than linear discriminant function (LDF) and quadratic discriminant function (QDF). However these unsupervised learning algorithms showed similar discriminatory power as supervised methods of discriminant analysis.

The aim of this study is to scrutinize and discuss the performance of four fundamental algorithms of discriminant analysis. Linear and quadratic discriminant analysis and their

*Corresponding author: serdarkuyuk@gmail.com
© China University of Geosciences and Springer-Verlag Berlin Heidelberg 2014

Manuscript received August 9, 2013.
Manuscript accepted November 29, 2013.

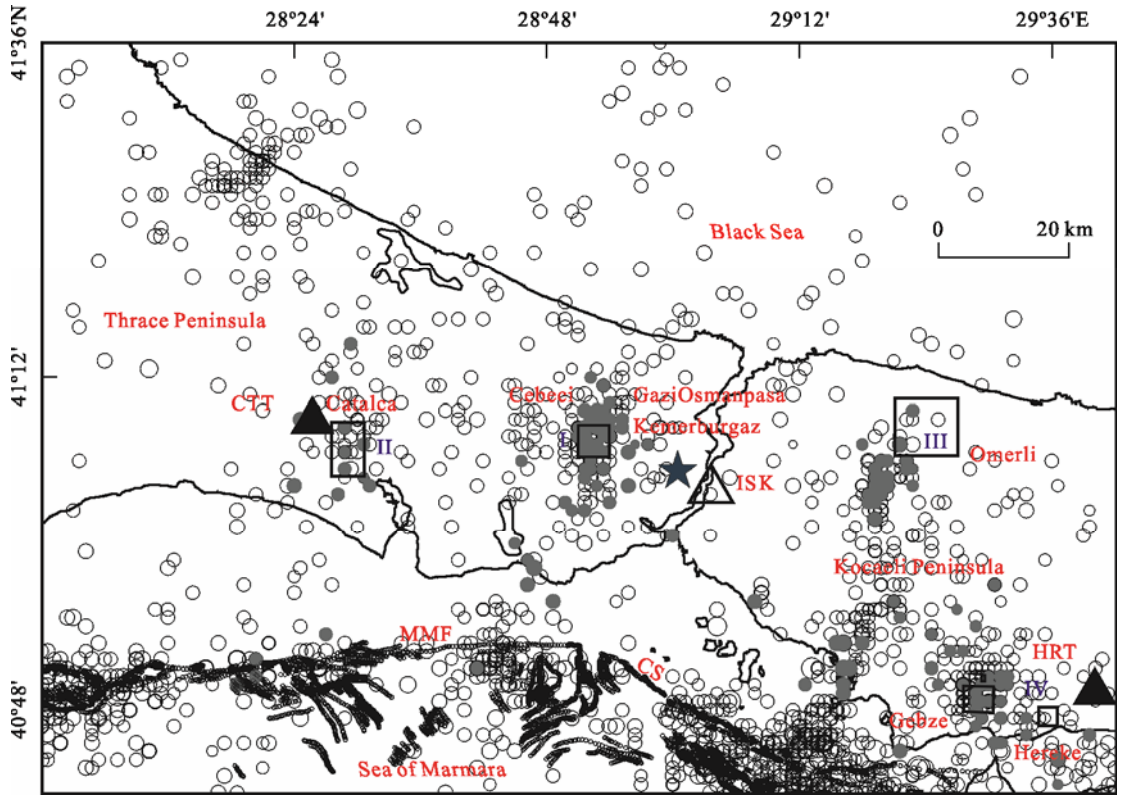

Figure 1. Map showing the study area and locations of seismic events used for statistical analysis of the seismicity catalogues (KOERI-NEMC) for 1995–2007 (first 8 months) marked by open black circles and locations of seismic events with duration magnitude between 1.8 and 3.0 (filled black circles) used for waveform of digital data for vertical seismograms recorded at ISK broad-band station (open triangle) and CTT and HRT short-period (filled triangles) stations (KOERI-NEMC, 2001–2004). Boxes show the quarry sites (I. Gaziosmanpaşa/Cebeci and Kemerburgaz; II. Çatalca; III. Ömerli; IV. Hereke) determined from satellite images (Musaoğlu et al., 2004) and field observations. Filled black star shows the location of 20th November 2005 explosion. MMF. Main Marmara fault; ÇS. Çinarcik segment of the MMF (adopted from Horasan et al., 2009).

derivative forms are evaluated. We have used two time variant parameters, complexity which is the ratio of integrated powers of the velocity seismogram, and S/P amplitude ratio as classifiers by using waveforms from 179 events ($1.8<M<3.0$). The performances of models are discussed for earthquakes and quarry blasts separately. Furthermore we proposed a new adjusted quadratic discrimination form which correctly classified 97.6% of events. Although errors of methods are less than 4%, we justified their confidence levels by cross validation and resubstitution errors.

2 DATA AND METHODS

The parameters S/P amplitude ratio and complexity are used for classification of seismic activities using statistical analysis. These parameters are obtained from Horasan et al. (2006) which is supported by Bogazici University Research Fund Project. The study area is located at 40.70ºN–41.60ºN latitude and 28.00ºE–29.70ºE longitude. The rectangles shown in Fig. 1 (I. Gaziosmanpaşa, II. Çatalca, III. Ömerli, IV. Gebze-Hereke) represent quarrying areas as determined by satellite and field observations (Horasan et al., 2006; Musaoğlu et al., 2004).

Horasan et al. (2009) pointed out that the seismic events with a magnitude less than three might be related to both earthquakes and man-made explosions from the quarries in the study area. No explosions with magnitude 3 or more observed in this study region. The number of seismic events in the quarries increased during the daytime interval of 7 a.m. to 4 p.m. GMT (9 a.m. and 6 p.m. local time) which corresponds to regular blasting hours of the quarries (Horasan et al., 2009). CTT and HRT have short-period seismometers while station ISK has a broad-band seismometer. The short-period seismometers for CTT and HRT stations were changed with broad-band one's after 2007. There are totally 179 events in the data set where 150 of them are QBs and 29 of them are EQs. Distribution of seismic events according to complexity and S/P ratio are shown in Fig. 2. Brown color asterisks show the complexity where green dots indicate the S/P ratios.

The first parameter S/P amplitude ratio was obtained from the P and S wave peak to peak amplitude measurements on the seismograms. The second parameter, complexity is the ratio of integrated powers of the velocity seismogram $s^2(t)$ in the selected time windows length (t_1 and t_2: first and second time window lengths; t_0: the onset time of P-wave). The complexity (C) can be expressed as follows (Arai and Yosida, 2004)

$$C: \int_{t_1}^{t_2} S^2(t)\mathrm{d}t \Big/ \int_{t_0}^{t_1} S^2(t)\mathrm{d}t \qquad (1)$$

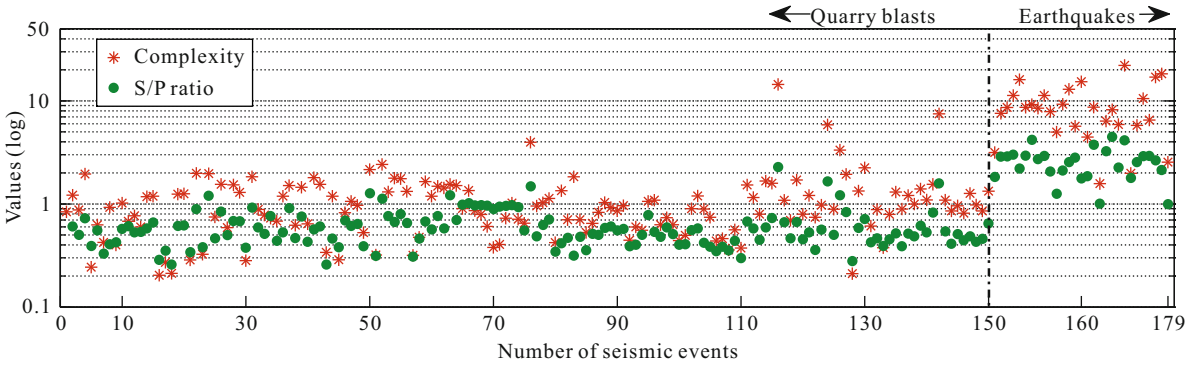

Figure 2. Distribution of seismic events according to complexity and S/P ratio. There are totally 179 events in the data set where 150 of them are QBs and 29 of them are EQs. Brown color shows the complexity which has higher values for earthquakes and green color indicates the S/P ratios.

3 DISCRIMINANT ANALYSIS

Discrimination analysis uses events to calculate the parameters of discriminant functions of the variables. This function could be whether linear or nonlinear. This separator determines the boundaries between classes. Linear discriminant analysis basically estimates one covariance matrix for all classes where quadratic discriminant analysis estimates one covariance matrix for each class. Diagonal linear or diagonal quadratic discriminants (DLDF, DQDF) uses only diagonal element of covariance matrices. Classification is achieved by minimizing the expected classification cost

$$\hat{y} = \underset{y=1,\ldots K}{\arg\min} \sum_{k=1}^{K} \hat{P}(k \mid x) C(y \mid k) \tag{2}$$

where $\hat{y}$ is the predicted classification, K is the number of classes, $\hat{P}(k \mid x)$ is the probability of class k for observation x, $C(y|k)$ is the cost of classifying an observation as y when its true class is k. Under this modeling assumption, fitting infers the mean and covariance parameters of each class and linear discriminant analysis computes the sample mean of each class. Then it computes the sample covariance by first subtracting the sample mean of each class from the observations of that class, and taking the empirical covariance matrix of the result (Matlab, 2012). The functional form is

$$f = \mathrm{m} + \mathbf{n}^{\mathrm{T}} \mathrm{X} \tag{3}$$

where $\mathrm{X} = \begin{bmatrix} \mathrm{C} \\ \mathrm{S/P} \end{bmatrix}$, m is constant and **n** is (a vector) coefficients of the equation.

On the other hand quadratic discriminant analysis computes the sample mean of each class. Then it computes the sample covariance by first subtracting the sample mean of each class from the observations of that class, and taking the empirical covariance matrix of each class (Matlab, 2012). The form of function is

$$f = \mathrm{m} + \mathbf{n}^{\mathrm{T}} \mathrm{X} + X^{T} \mathrm{E} X \tag{4}$$

where E is the matrix of quadratic coefficients. The fit of both methods does not use any prior probabilities or costs for fitting. As a result, below equation defines the classes

$$F = \begin{cases} EQ & f < 0 \\ QB & f \geq 0 \end{cases} \tag{5}$$

4 RESULTS AND DISCUSSIONS

Various studies indicate that time dependent parameters, the complexity and S/P amplitude ratio could be used as a criteria to discriminate man-made explosions from earthquakes (Yildirim et al., 2011; Kuyuk et al., 2010; Horasan et al., 2009). However it is a fact that clustering algorithms which might be competent in one region could be incompetent in other regions due to local site and source effects, geological structure of path etc. (Zeiler and Velasco, 2009).

A conventional, discriminant analysis which uses input events to estimate the parameters of discriminant functions of the predictor variables are applied. Discriminant functions resolve the boundaries, in predictor space, between two classes. The resulting classifier then discriminates among the classes based on the predictors.

Decision boundaries of clustering which show the optimal trade-off between performance on the data set and simplicity of classifier, thereby giving the highest accuracy using different algorithms LDF, QDF, DLDF, DQDF, are shown in Fig. 3. Red reverse triangles and black circles indicate earthquakes and quarry blasts, respectively. Discriminant functions for each methods is given below

$$f_{LDF} = -18.47 + \begin{bmatrix} 0.56 & 8.77 \end{bmatrix} \begin{bmatrix} C \\ SP \end{bmatrix} \tag{6}$$

$$f_{QDF} = -5.99 + \begin{bmatrix} -1.61 & -9.73 \end{bmatrix} \begin{bmatrix} C \\ SP \end{bmatrix} + \begin{bmatrix} C & SP \end{bmatrix} \begin{bmatrix} 0.63 & -2.40 \\ -2.40 & 14.41 \end{bmatrix} \begin{bmatrix} C \\ SP \end{bmatrix} \tag{7}$$

$$f_{DLDF} = -25.17 + \begin{bmatrix} 1.39 & 10.32 \end{bmatrix} \begin{bmatrix} C \\ SP \end{bmatrix} \tag{8}$$

$$f_{DQDF} = -7.42 + \begin{bmatrix} -0.25 & -3.91 \end{bmatrix} \begin{bmatrix} C \\ SP \end{bmatrix} + \begin{bmatrix} C & SP \end{bmatrix} \begin{bmatrix} 0.28 & 0 \\ 0 & -3.91 \end{bmatrix} \begin{bmatrix} C \\ SP \end{bmatrix} \tag{9}$$

The overall accuracies are 96.6%, 96.6%, 95.5% and 96.6% (see Table 1). LDF, QDF and DQDF have the same accuracy with different boundaries. Although quadratic discrimanant analysis fits multivariate normal densities with the estimation of covariance matrix where diaquadratic discrimanant analysis fits with the estimation of just diagonal covariance matrix, the total accuracy is same with different error rates by two classes.

Quadratic discriminant analysis resembles LDF, where it is assumed that there are only two classes of points, and that the measurements are normally distributed. Dissimilar to LDF however, in QDF there is no supposition that the covariance of each of the classes are equal. LDF has the least misclassified estimation with one error for quarry blasts. However this is not valid for earthquakes because it gave the worst estimation with 82.8% accuracy with the same error rate of DLDF (Table 1).

4.1 Resubstitution Error and Confusion Matrix

Resubstitution errors is the error rate of the training data. It is the error estimate based on the difference between the predicted values of trained model and the observed values in the training dataset. If the resubstitution error is low, the prediction of the classifier is considered to be good. However, having low resubstitution error does not assure good predictions for new data. Resubstitution error is often an overly optimistic estimate of the predictive error on new data (Matlab, 2012). Estimated resubtitution errors are shown in Table 2. All methods are satisfactorily well where LDF, QDF and DQDF errors 75% lower than DLDF.

Moreover, the confusion matrix indicates how many errors, and which types, arise in resubstitution. Because this problem has 2 classes, the confusion matrix C is a 2-by-2 matrix. Confusion matrices are given below

$$C_{LDF} = \begin{bmatrix} 149 & 1 \\ 5 & 24 \end{bmatrix} \tag{10}$$

$$C_{QDF} = \begin{bmatrix} 147 & 3 \\ 3 & 26 \end{bmatrix} \tag{11}$$

$$C_{DLDF} = \begin{bmatrix} 147 & 3 \\ 5 & 24 \end{bmatrix} \tag{12}$$

$$C_{DQDF} = \begin{bmatrix} 146 & 4 \\ 2 & 27 \end{bmatrix} \tag{13}$$

Confusion matrix of LDF indicates that 149 quarry blasts out of 150 are classified correctly, and one quarry blast misclassified as earthquake whereas 24 earthquakes are classified correctly and 5 earthquakes are misclassified as quarry blast. These interpretations can be extended to other confusion matrices where total results are shown in Table 1 with the accuracy percentages of each method.

4.2 Cross Validation

Typically, discriminant analysis classifiers are robust and do not show overtraining when the number of predictors is much less than the number of observations. Nevertheless, it is good practice to cross validate your classifier to ensure its stability.

We have divided five-fold cross validation of all discriminant analysis classifiers and cross-validation loss for models, meaning the error of the out-of-fold observations. This method uses models trained on in-fold observations to predict response for out-of-fold observations. Every training fold contains roughly 4/5 of the data and every test fold contains roughly 1/5 of the data. The first model was trained on Complexities and S/P ratios and correspondence labels with the first 1/5 excluded, the second model was trained on the same input and output with the second 1/5 excluded, and so on. Then, we computed predictions for the first 1/5 of the data using the first model, for the second 1/5 of data using the second model, and so on (Matlab, 2012).

The cross-validated loss is nearly as low as the original resubstitution loss. Therefore, we have confidence that the classifier is reasonably accurate. On the other hand, LDF and QDF has lower errors than diagonal LDF and QDF which means usage of just diagonals makes confidence measures lower.

Table 1 Comparisons of the methods

Methods	# of events	# of QB	Misclassified QB	% of accuracy for QB	# of EQs	Misclassified EQ	% of accuracy for EQ	% of total accuracy
LDF	179	150	1	99.3	29	5	82.8	96.6
QDF	179	150	3	98.0	29	3	89.7	96.6
DLDF	179	150	3	98.0	29	5	82.8	95.5
DQDF	179	150	4	97.3	29	2	93.1	96.6
QDF*	179	150	1	99.3	29	3	89.7	97.6

*. Adjusted quadratic discriminant function (Eq. 14).

Table 2 Comparisons of the accuracy and errors of the methods

Methods	Resubtitution error	Cross validation errors	# of total misclassified events	% of total accuracy
LDF	0.033 5	0.033 5	6	96.6
QDF	0.033 5	0.033 5	6	96.6
DLDF	0.044 7	0.039 1	8	95.5
DQDF	0.033 5	0.039 1	6	96.6

4.3 Adjusted Quadratic Discriminant Function

We further updated the quadratic function by adding two additional parameters, a, b to constant and quadratic terms of the function. By doing so we tried to shift quadratic function in order to contain two additional quarry blasts by not losing

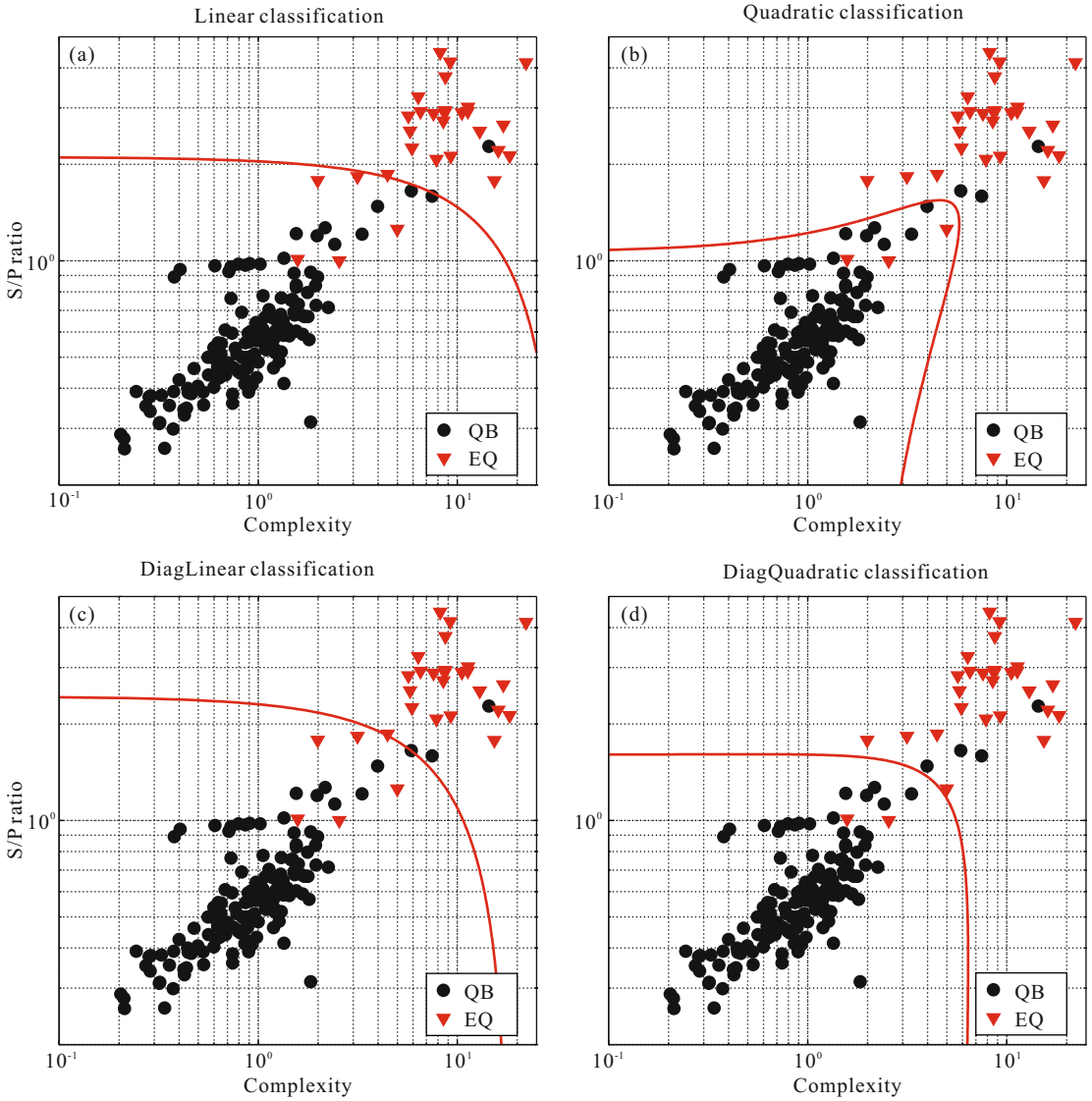

Figure 3. Classification with discriminant analysis, decision boundaries (red lines) represent the optimal trade-off between performance on the data set and simplicity of classifier. Black circles and red triangles indicate quarry blasts and earthquakes. Four different algorithms (a) linear discriminant function, (b) quadratic discriminant functions, (c) diaglinear discriminant functions, (d) diagquadratic discriminant functions are shown. Except DLDF with 8 missclassification, other algorithm misclassified 6 events with different number of each class. Linear and Quadratic discriminant analysis fits multivariate normal densities with the estimation of covariance matrix where diaglinear and diagquadratic discriminant analysis fits with the estimation of diagonal covariance matrix.

any miscategorized earthquakes (Fig. 4). This arbitrary adjustment gave the best possible accuracy, 97.6% by total 4 misclassifications (one quarry blast and 3 earthquakes, Table 1). The updated function is given as

$$f = -(5.99 + a) + \begin{bmatrix} -1.61 & -9.73 \end{bmatrix} \begin{bmatrix} C \\ SP \end{bmatrix} + \begin{bmatrix} C & SP \end{bmatrix} \begin{bmatrix} 0.63 & -(2.40 + b/2) \\ -(2.40 + b/2) & 14.41 \end{bmatrix} \begin{bmatrix} C \\ SP \end{bmatrix} \quad (14)$$

where a=1.4 and b=3.7 with the same coefficients rest.

Even though this alteration might gave the best result, this might be a particular solution and equation might be data and region dependant rather than a general solution. Nevertheless, clustering seismic events are vastly dependant on area, data, and methodology.

Horasan et al. (2009) used LDF on S/P ratio with log$S$ and complexity with Sr (spectral ratio) together to discriminate the earthquakes and the quarry blasts in the same region. They obtained similar results using LDF analysis from the amplitude ratio with log$S$. They examined the region by separating the events location in to the four locations. On the contrary, this study evaluated all events at once and the predictors implemented using whole data. This study is giving a much general image and these clustering techniques are more accurate for the vicinity of Istanbul. Kuyuk et al. (2012, 2011) also applied self-organizing map, k-means and Gaussian mixture models for the same data set with the 94.0%, 95.0%, and 96.1% of accuracy. Adjusted QDF in the present study achieved better success where applications of algorithm are rather simple and faster than other even

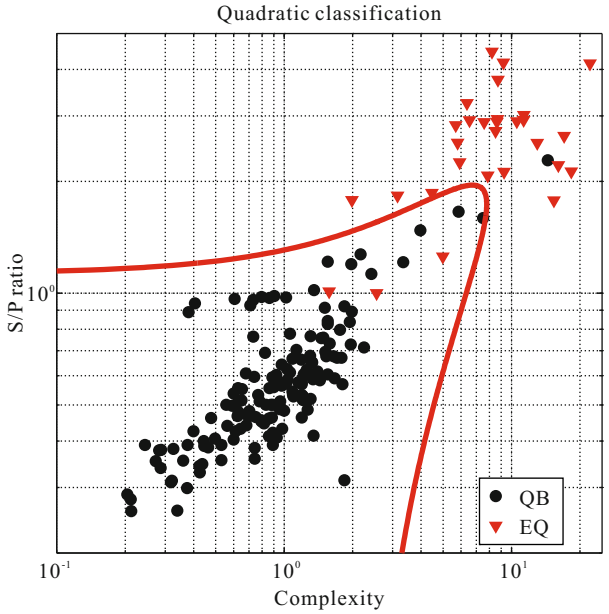

Figure 4. Adjusted quadratic function by adding two additional parameters, quadratic function shifted to contain additional quarry blasts by not losing any miscategorized earthquakes. This arbitrary adjustment gave the best possible accuracy, 97.6% by total 4 misclassifications (one quarry blast and 3 earthquakes, Table 1).

complex clustering methods. However three algorithms used by Kuyuk et al. (2012, 2011) were unsupervised algorithms that no a-priori information was fed to train methods.

As a result, adjusted QDF improved the result of QDF and LDF. On the other hand, it should be emphasized that LDF is computationally effortless where QDF is proposing more complex equation. Table 1 summarizes all the results together with the misclassified quarry blasts and earthquakes.

REFERENCES CITED

Arai, N., Yosida, Y., 2004. Discrimination by Short-Period Seismograms. International Institute of Seismology and Earthquake Engineering, Building Research Institute (IISEE). Lecture Note, Global Course, Tsukuba, Japan. 10

Bennett, T. J., Murphy, J. R., 1986. Analysis of Seismic Discrimination Capabilities Using Regional Data from Western United States Events. *Bulletin Seismological Society of America*, 76(4): 1069–1086

Del Pezzo, E., Esposito, A., Giudicepietro, F., et al., 2003. Discrimination of Earthquakes and Underwater Explosions Using Neural Networks. *Bulletin of the Seismological Society of America*, 93(1): 215–223

Gitterman, Y., Pinky, V., Shapira, A., 1998. Spectral Classification Methods in Monitoring Small Local Events by the Israel Seismic Network. *Journal of Seismology*, 2(3): 237–256

Horasan, G., Bozepe-Güney, A., Küsmezer, A., et al., 2006. İstanbul Ve Civarindaki Deprem Ve Patlatma Verilerinin Birbirinden Ayirt Edilmesi Ve Kataoglanmasi (Discrimination and Cataloging of Quarry Blasts and Earthquakes in the Vicinity of Istanbul). Report 05T202, Boğaziçi University Research Foundation Bebek-İstanbul. 76

Horasan, G., Boztepe-Güney, A., Küsmezer, A., et al., 2009. Contamination of Seismicity Catalogs by Quarry Blasts: An Example from Istanbul and Its Vicinity, Northwestern Turkey. *Journal of Asian Earth Sciences*, 34(1): 90–99

Jenkins, R. D., Sereno, T. S., 2001. Calibration of Regional S/P Amplitude-Ratio Discriminants. *Pure Appl. Geophys*., 158(7): 1279–1300, doi:10.1007/PL00001223

Koch, K., Fah, D., 2002. Identification of Earthquakes and Explosions Using Amplitude Ratios: The Vogtland Area Revisited. *Pure Appl. Geophys*., 159(1): 735–757

Kuyuk, H. S., Yildirim, E., Dogan, E., et al., 2010. Self Organizing Map Approach for Discrimination of Seismic Event and Quarry Blasts in the Vicinity of Istanbul. 14th European Conference on Earthquake Engineering, Ohrid, Republic of Macedonia

Kuyuk, H. S., Yildirim, E., Dogan, E., et al., 2011. An Unsupervised Learning Methodology: Application to Discrimination of Seismic Event and Quarry Blasts in the Vicinity of Istanbul. *Natural Hazards and Earth Science*, 11(1): 93–100

Kuyuk, H. S., Yildirim, E., Dogan, E., et al., 2012. Application of *k*-Means and Gaussian Mixture Model for Classification of Seismic Activities in Istanbul. *Nonlinear Processes in Geophysics*, 19(4): 401–419

Matlab, 2012. Statistical Toolbox (vR2011b)

Musaoğlu, N., Coşkun, M. Z., Göksel, Ç., et al., 2004. Istanbul Anadolu Yakasi Hazine Arazilerinin Uydu Verileri ve Coğrafya Bilgi Sistemleri (CBS) Ile Incelenmesi (Investigation of State Owned Lands in Anatolian Side of Istanbul by Satellite Data and Geographic Information System). TÜBITAK, Project No: 102I022 (IÇTAG-I 433). 69

Musil, M., Plešinger, A., 1996. Discrimination between Local Microearthquakes and Quarry Blasts by Multi-Layer Perceptrons and Kohonen Maps. *Bulletin of the Seismological Society of America*, 86(4): 1077–1090

Rodgers, A. J., Walter, W. R., 2002. Seismic Discrimination of the May 11, 1998 Indian Nuclear Test with Short-Period Regional Data from Station NIL (Nilore, Pakistan). *Pure Appl. Geophys*., 159: 679–700

Ursino, A., Langer, H., Scarfi, L., et al., 2001. Discrimination of Quarry Blasts from Tectonic Microearthquakes in the Hyblean Plateau (Southeastern Sicily). *Annali di Geofisica*, 44(4): 703–722

Wüster, J., 1993. Discrimination of Chemical Explosions and Earthquakes in Central Europe—A Case Study. *Bulletin of the Seismological Society of America*, 83(4): 1184–1212

Yildirim, E., Gulbag, A., Horasan, G., et al., 2011. Discrimination of Quarry Blasts and Earthquakes in the Vicinity of Istanbul Using Soft Computing Techniques. *Comput. Geosci*., 37(9): 1209–1217

Zeiler, C., Velasco, A. A., 2009. Developing Local to Near-Regional Explosion and Earthquake Discriminants. *Bulletin of the Seismological Society of America*, 99(1): 24–35