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

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

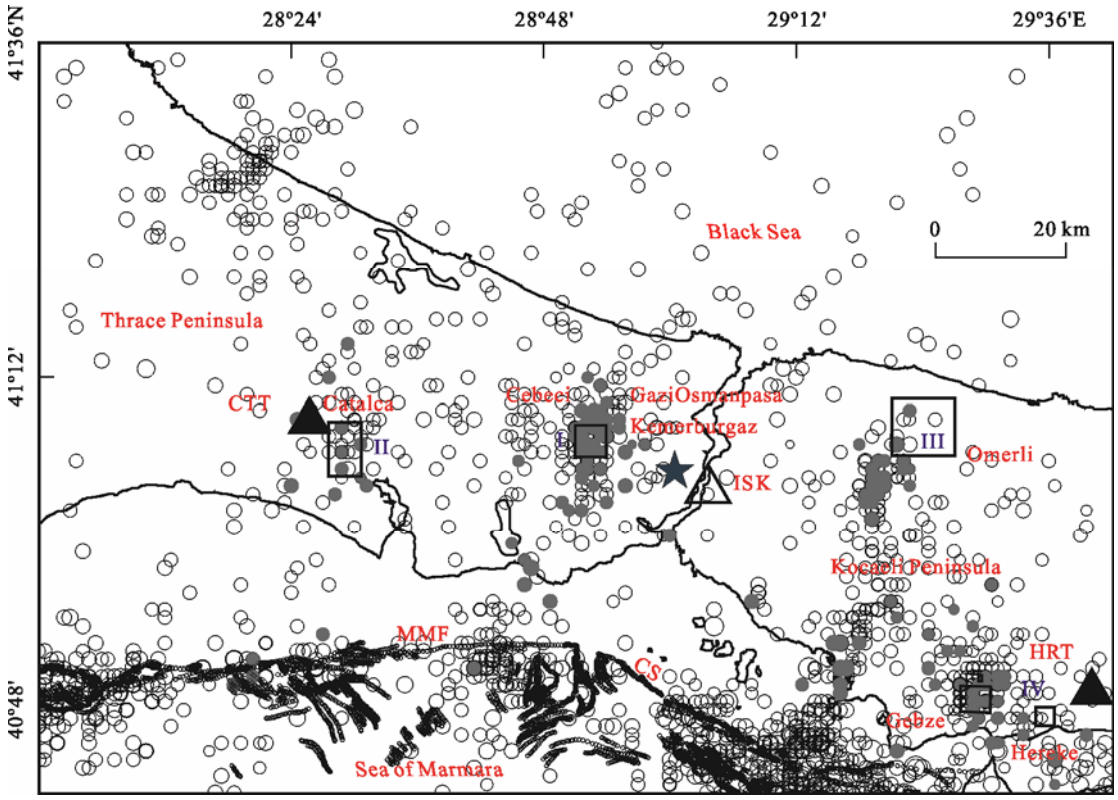

Figure 1. Map showing the study area and locations of seismic events used for statistical analysis of the seismicity catalogues (KOERI-NEMC) for 1995–2007 (first 8 months) marked by open black circles and locations of seismic events with duration magnitude between 1.8 and 3.0 (filled black circles) used 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)

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