An unsupervised learning algorithm: application to the discrimination of seismic events and quarry blasts in the vicinity of Istanbul

H. S. Kuyuk\textsuperscript{1}, E. Yildirim\textsuperscript{2}, E. Dogan\textsuperscript{1}, and G. Horasan\textsuperscript{2}

\textsuperscript{1}Department of Civil Engineering, Sakarya University, Sakarya, Turkey
\textsuperscript{2}Department of Geophysical Engineering, Sakarya University, Sakarya, Turkey

Received: 21 June 2010 – Revised: 3 December 2010 – Accepted: 3 December 2010 – Published: 8 January 2011

Abstract. The results of the application of an unsupervised learning (neural network) approach comprising a Self Organizing Map (SOM), to distinguish micro-earthquakes from quarry blasts in the vicinity of Istanbul, Turkey, are presented and discussed. The SOM is constructed as a neural classifier and complementary reliability estimator to distinguish seismic events, and was employed for varying map sizes. Input parameters consisting of frequency and time domain data (complexity, spectral ratio, S/P wave amplitude peak ratio and origin time of events) extracted from the vertical components of digital seismograms were estimated as discriminants for 179 (1.8 < \textit{M}_d < 3.0) local events. The results show that complexity and amplitude peak ratio parameters of the observed velocity seismogram may suffice for a reliable discrimination, while origin time and spectral ratio were found to be fuzzy and misleading classifiers for this problem. The SOM discussed here achieved a discrimination reliability that could be employed routinely in observatory practice; however, about 6\% of all events were classified as ambiguous cases. This approach was developed independently for this particular classification, but it could be applied to different earthquake regions.

1 Introduction

Neural networks (NN) are one of the popular recent mathematical/computational alternative techniques that have been applied to seismic event classification problems (Taylor et al., 1989; Falsaperla et al., 1996; Tiira, 1999; Del Pezzo et al., 2003; Scarpetta et al., 2005; Yildirim and Horasan, 2008; Kuyuk et al., 2009). NN learning algorithms can be categorized into three groups with respect to the type of feedback to which the learner has access. These are: (1) supervised learning, (2) reinforcement learning and (3) unsupervised learning. Unsupervised learning is distinguished from the others in that the learner is given only unlabeled examples. Unsupervised learning models include factor analysis, principal component analysis (PCA), mixtures of Gaussians, independent component analysis (ICA), hidden Markov models, state-space models (Ghahramani, 2004), adaptive resonance theory (ART) and many variants and extensions, but the SOM is the most commonly used algorithm. In particular, for large-scale high-dimensional data sets, a SOM allows a sensitive visualization of the data by vector quantization and dimension reduction. Based on the relatively simple SOM representation, further processing, such as clustering or feature grouping, can be achieved. Numerous methods exist in the literature for the discrimination of different typologies of seismic events (Gitterman et al., 1999; Joswig, 1990). In the seismological context, recent studies have reported that the SOM may offer a promising alternative classification method (Musil and Plesinger, 1996; Allamehzadeh and Mokhtari, 2003; Esposito et al., 2007).

In this study, the SOM method is applied to the discrimination of earthquakes (EQs) and quarry blasts (QB) recorded by three seismic stations (ISK, CTT, HRT) in Istanbul and its vicinity. The research area is located in the Marmara region, northwest Turkey. The local MARNET network, which is operated by the Kandilli Observatory and Earthquake Research Institute (KOERI) in order to monitor and record the earthquakes in the Marmara Region, was founded in 1976. These stations do not only record earthquakes, but also record quarry blasts.
The tectonic processes forming the Sea of Marmara and its surrounding area are controlled by the North Anatolian Fault Zone (NAFZ). NAFZ is a dextral strike-slip fault zone, extending from the Karliova region to the Gulf of Izmit along Anatolia, and south of Thrace as the Ganos Fault (Şengör et al., 1985; Barka and Kadinsky-Cade, 1988). Many micro and macro tectonic earthquakes have occurred due to the NAFZ and its segments in the vicinity of Istanbul. A few areas of quarrying are in operation in Istanbul and its near surroundings: Gaziosmanpasa (Cebeci and Kemerburgaz), Çatalca, Omerli, Gebze-Hereke (Musaoglu et al., 2004). Quarrying areas are illustrated in Fig. 1.

Currently, quarrying regularly contaminates the seismograms recorded by the KOERI, National Earthquake Monitoring Center (NEMC) seismic network. Accurate discrimination between explosions in quarries and earthquakes is important for analyzing active tectonic and seismic risk of the region. Recorded explosions in the catalogues can mislead scientists interpreting the active tectonics and can lead to erroneous results in the analysis of seismic hazards in the study area (Gökşan et al., 2002). Therefore, it is important to distinguish QBs from EQs in the KOERI NEMC seismicity catalog.

The aim of this study was to apply one of the unsupervised learning methods (in this case, SOM) for discriminating QBs from tectonic EQs. It was found that the SOM technique could be employed successfully using four discriminants for classifying seismic events such as micro earthquakes and quarry blasts.

2 Data and methods

Investigations performed by Horasan et al. (2006) around Istanbul revealed that the number of events with magnitude $M_d < 3$ between 1995 and 2007 was affected by the following two factors: (1) contamination of the seismicity catalogs with blasts regularly created in quarries at certain time intervals every day, and (2) the existence of imprecisely reported seismic events with duration magnitude less than 3 in the seismicity catalogs (Horasan et al., 2009). From the results of preliminary work, three factors needed to be taken into consideration when deciding which methods are to be used for the discrimination of QBs from EQs in the study area.

First, utilization of satellite images may not be accurate since the locations of quarries, especially in the western part of the study area, are along the recently defined active fault zone (Fig. 1) suggested by Gökşan et al. (2002). Second, utilization of the statistical distribution of seismic events during daytime and night time hours (Wiemer and Baer, 2000) may not be accurate since the seismic events with magnitude less than 3 were not precisely reported by KOERI NEMC for the study area. Third, information is missing regarding the locations of QBs, blasting times and the number of blasts from quarries, since these data have not been reported to KOERI, NEMC on a regular basis; a completely quarry blast-free seismicity catalog is therefore not possible.
In this study, four parameters widely used in literature (amplitude peak ratio, power ratio, spectral amplitude ratio and origin time) are considered (Pomeroy et al., 1982; Bennett and Murphy, 1986; Baumgardt and Young, 1990; Wüster, 1993; Gitterman et al., 1998; Wiemer and Baer, 2000). The vertical components of the digital velocity seismograms (recorded by Istanbul, ISK, Catalca, CTT, and Hereke, HRT, stations) of seismic events with duration magnitude ($M_d$) between 1.8 and 3.0 (KOERI, NEMC) that occurred between 2001 and 2004 are used. Out of a total of 179 seismic events, 16 were recorded at Catalca, 71 at Gaziosmanpasa, 48 at Hereke and 44 at Omerli (Fig. 1). The data (complexity, spectral ratio, S/P amplitude peak ratio) used in this study were obtained from Horasan et al. (2006) with the support of the Bogazici University Research Fund (Project No: 05T202).

The four parameters investigated in the analysis are defined as follows:

1. Complexity ($C$) is the ratio of integrated powers of the seismogram in the selected time windows ($t_1 = 2$ s; $t_2 = 4$ s; $t_0$ is the onset time of the P-wave). $C$ can be expressed as follows (Arai and Yosida, 2004):

$$C = \frac{\int_{t_1}^{t_2} a^2(t) dt}{\int_{t_0}^{t_1} a^2(t) dt}$$ (1)

The limits of the integrals of $C$ given in Eq. (1) were determined by a trial-and-error approach to find the best representative $C$ values for both blasts and earthquakes. The integrals were computed separately for the ISK, CTT and HRT seismic stations. The complexity has a higher value for earthquakes because the S-wave amplitude of the earthquake waveform is greater than the P-wave amplitude.

2. Spectral ratio ($Sr$) is the ratio of integrated spectral amplitudes $a(w)$ of the seismogram in the selected frequency bands (high frequency band, $f_1 - f_2$: 5–10 Hz; low frequency band, $f_0 - f_1$: 1–5 Hz). $Sr$ can be written as (Gitterman and Shapira, 1993):

$$Sr = \frac{\int_{f_1}^{f_2} a(w) dw}{\int_{f_0}^{f_1} a(w) dw}$$ (2)

The limits of integrals ($f_0$, $f_1$, $f_2$) were determined by comparing the spectra of quarry blasts with those of earthquakes.

3. S/P ratio is the relative ratio of the amplitudes of S-waves to those of P-waves. This parameter was obtained from the P- and S-wave peak to peak amplitude measurements of the seismograms using GURALP visualization program Scream 4.3. The peak ratio of S-to P-waves is smaller for quarry blasts, since the S-wave amplitude of the seismogram is smaller than the P-wave amplitude.

4. Time refers to the origin time of an event. The number of seismic events in the quarries increased during the day (7 a.m.–4 p.m. GMT, or 9 a.m.–6 p.m. LT), coincident with regular blasting hours of the quarries (Horasan et al., 2009).

Table 1 shows statistical values of these parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity</td>
<td>2.4413</td>
<td>0.2040</td>
<td>22.1300</td>
<td>3.7338</td>
</tr>
<tr>
<td>Spectral Ratio</td>
<td>1.9191</td>
<td>0.1500</td>
<td>9.3000</td>
<td>1.3643</td>
</tr>
<tr>
<td>S/P</td>
<td>0.9300</td>
<td>0.0600</td>
<td>4.4660</td>
<td>0.8482</td>
</tr>
<tr>
<td>Time</td>
<td>12.2849</td>
<td>1.0000</td>
<td>23.0000</td>
<td>3.8457</td>
</tr>
</tbody>
</table>

### 2.1 Self Organizing Map (SOM)

A SOM is in the competitive Artificial Neural Networks (ANNs) category, has no hidden layer, and contains no nonlinear activation function. Therefore it is strictly linear in its response (Kohonen, 1990). A SOM carries out unsupervised learning with sets of unknown categorized data and discovers classes by clustering inputs according to various similarity criteria. Namely, competitive ANNs are capable of discerning the common features or topology of different input vectors without any subsidiary information. Thus, methods like clustering or feature grouping can be carried out easily using a SOM representation.

The SOM algorithm uses Euclidian metrics to measure distances between vectors; thus, before training, input data are firstly linearly normalized. Linear initialization and a batch training algorithm are performed. Regression of an ordered set of model vectors $\varphi_i$ into the space of observation vectors $k$ is executed by the following process:

$$\varphi_i(t+1) = \varphi_i(t) + \Gamma_{c(t), i}(k(t) - \varphi_i(t))$$ (3)

where $t$ is the step index, whereby the regression is performed recursively for each presentation of a sample of $k$. Index $c$ (“winner”) is defined by the condition

$$\| k(t) - \varphi_c(t) \| \leq \| k(t) - \varphi_i(t) \|$$ (4)

where $\Gamma_{c(t), i}$ is the neighborhood function, a kind of smoothing kernel that is time-dependent and with a location dependent on the condition in Eq. (3). It is a minimizing function of the distance between the $i$-th and $c$-th models.

Table 1. Fundamental descriptive statistical information about the data and their ranges.
Fig. 2. SOM Visualization of predictions and data (7 × 7 map size). Component planes: (a) complexity, (b) spectral ratio (Sr), (c) S/P ratio, (d) time and (e) map unit labels. In Fig. 1e QB and EQ represent quarry blasts and earthquakes, respectively. Unlabeled cells could not be clustered by the method. Five figures have connection by position; the hexagon in a certain position corresponds to the same map unit in each figure. The lightness of the colors increases with indicator value. Two components, complexity and S/P ratio have a strong correlation with the label node matrix.

\[
\Gamma_{c(t),i} = \alpha(t) \exp\left(-\frac{||k_i - k_c||^2}{2\eta^2(t)}\right)
\]

where \(\alpha(t)\) is the learning rate, which decreases with each step, and is between 0 and 1. \(k_i, k_c\) are the vectorial locations in the display grid and \(\eta(t)\) corresponds to the width of the neighborhood function (Kohonen, 1997; Kohonen et al., 1996). A MATLAB\textsuperscript{1} based SOM Toolbox developed by the SOM Toolbox team, Laboratory of Computer and Information Science (Helsinki University of Technology, Finland) was utilized for analysis. Further information can be obtained from the technical manual of the toolbox (Technical manual, 2000).

3 Results and discussion

Parameters derived from time- and frequency-domain analyses of the seismograms (amplitude peak ratio, power ratio, and spectral amplitude ratio) were used for discrimination. Discriminative parameters are illustrated in Fig. 2, where component planes and corresponding map unit labels are provided. The labels in hexagons of the label matrix show the corresponding neuron. The units are very close to their neighbors in the bottom-left corners of complexity, Sr and S/P ratio. A border can also be seen between the top 2 neurons and the remainder for complexity and S/P ratio. The map unit in the top-right corner has high values for the same two parameters. Moreover, high values in the top left corners were observed for time and for spectral ratio. Visually, there is a strong correlation between the classification of events with complexity and with S/P ratio.

The projection of the parameter set onto the subspace spanned by its two eigenvectors with greatest eigenvalues is shown in Fig. 3 for different map-sizes. The two labels have been plotted such that red stars indicate earthquakes and blue asterisks represent quarry blasts. The SOM grid has been projected onto the same subspace. Neighboring map units are connected with gray lines. Labels associated with map units are also shown on nodes (neurons) as EQs and QBs. The distribution of the data set is compatible with the SOM. Figure 3a, b, c, d, e and f corresponds to 6 × 6, 8 × 8, 10 × 10, 12 × 12, 14 × 14 and 20 × 20 map sizes, respectively. As the map-size is enlarged, the SOM algorithm adapts to the distribution of the input data while the the number of nodes without labels increases. Node density is boosted around events with increments of map-size.

The four variables represented by bar charts for each map unit (hexagons) in Fig. 4 are complexity, spectral ratio, S/P ratio and time, respectively. For better representation, each parameter was normalized by dividing by the maximum value of each set. Color indicates the classes: red, green and

---

\textsuperscript{1}MATLAB software, ©1994–2010 The MathWorks, Inc.
Fig. 3. SOM and projection of the seismic events onto the subspace spanned by its two eigenvectors with greatest eigenvalues for different map-sizes. Red stars indicate earthquakes (EQs) and blue asterisks represent quarry blasts (QB). Neighboring map units are connected with gray lines. Labels associated with map units are also shown on nodes (neurons). Parts (a), (b), (c), (d), (e), (f) correspond to $6 \times 6$, $8 \times 8$, $10 \times 10$, $12 \times 12$, $14 \times 14$ and $20 \times 20$ map-sizes, respectively. As the map size is enlarged, the SOM algorithm adapts to distribution of the input data.
Fig. 4. Clustering results of the $7 \times 7$ dimension SOM. Red, green and black hexagons indicate earthquakes, quarry blasts and unlabeled events. The bar charts in each map unit show the normalized values of complexity, spectral ratio, S/P and time, respectively. Bar chart values are correlated with event discrimination. Values of the first and third bars, which indicate complexity and S/P ratios, respectively, in red cells (earthquakes) are higher than corresponding bars in green cells (quarry blast).

Fig. 5. Red stars indicate earthquakes and blue asterisks represent the quarry blasts; gray lines show the connections between neighboring map units. Red, blue and black circles located at the nodes indicate earthquakes, quarry blast and unlabeled events, respectively, and are classification results from the SOM. (a) Scatter plot of four discriminants in four dimensions; marker size represents the forth component, time. Large markers indicate late seismic activity. Parts (b), (c), (d) are the projections onto the corresponding parameters (axes are log scale).

black hexagons indicate EQs, QBs and ambiguous events. There were 49 nodes in the SOM. Three nodes could not be distinguished by the algorithm, so the model yielded a success rate of about 94%. Complexity and S/P ratio (first and third bars) were higher for EQs, and lower for QBs.

Projections of the 3-D plot illustrated in the top left are given in Fig. 5b, c, and d (logarithmic scale). The three coordinates represent the first three variables and the size of the marker shows the size of the fourth variable, time. Red stars indicate EQs and blue stars indicate QBs. Circles represent the map units (neurons) of the SOM. Nodes are classified into two classes after training. Black circles
indicate that the SOM could not decide whether the node belongs to EQs or QBs. Earthquakes occurring in the evenings are shown as larger stars. However, there was no direct correlation between time and event discrimination. Spectral ratio is a fuzzy parameter for this event as there was no correlation for high/small values in Fig. 5b. However, both spectral ratio and complexity were better discriminants, as there was a linear link as shown in Fig. 5c and d.

The SOM investigated here achieved a discrimination reliability that could be employed in observatory practice; however, about 6% of all events were classified as ambiguous cases. The accuracy might improve with other preprocessing strategies of data such as Linear Prediction Coding, wavelet techniques, etc. (Esposito et al., 2006).

4 Conclusions

The presence of quarries in an active seismic zone can cause errors during the analysis of the distribution of micro seismicity and the editing of seismic catalogs. This study demonstrated that the SOM technique is able to discriminate EQs from QBs using relevant data sets in the study area. The SOM algorithm proposes a model that optimally describes the domain of (discrete or continuously distributed) observations. The model is organized into a consequential two-dimensional order in which related models are closer to each other in the grid than the more different models. Thus the SOM is a kind of relationship graph, or a clustering diagram. Its computation is a nonparametric, recursive regression process. The data sets computed from the vertical component of waveform (including EQs and QBs) were recorded in Istanbul and its vicinity between 2001 and 2004 at three stations. From the results obtained, the proposed SOM model appears to be an efficient tool for distinguishing seismic events.

Acknowledgements. We thank the Bogazici University Research Fund and members of project no. 05T202. We thank Masato Motosaka for allowing us to use MATLAB Software\(^1\). We also thank the referees for their valuable comments.

Edited by: M. E. Contadakis
Reviewed by: three anonymous referees

References


