UNSUPERVISED SPECTRAL PATTERN RECOGNITION BY SELF ORGANIZING MAPS (SOM) ANALYSIS FOR SITE EFFECTS ESTIMATION: A REVIEW

R. Carniel 1, L. Barbui 1,2

1 Dipartimento DIEA, Università degli Studi di Udine, Italy
2 Dipartimento DIGE, Università degli Studi di Udine, Italy

This work aims to present a few enhancements to our previous work (Carniel et al., 2009) that introduced a post-processing application of a kind of artificial neural network called Self-Organizing Map (Kohonen, 1982) to improve the HVSR technique (Nakamura, 1989, 2000). Applying the SOM procedure at several seismic tremor traces some problems and considerations have been arisen about data pre-processing, SOM learning process and SOM results interpretation.

Using non-invasive geophysical techniques, data interpretation could be problematic both when it’s necessary to analyse large and multidimensional data sample and when the data are a few and highly noisy (Carniel et al., 2006, 2008; Barazza et al., 2009). In the first case SOM allows the unsupervised pattern recognition to project data onto a low dimensional map space (clustering), allowing an overview of the input space structure. Furthermore the method is completely generic, independent from the specific data type and also suitable to improve in-time analysis (Klose, 2006).

In the second case the issue is the opposite: it’s necessary to extract as much information as possible, then data pre-processing increases its importance. In order to estimate the amplitude spectrum, the classic FFT algorithm has been replaced by the Welch’s method (Welch, 1967) implementation (Lees, 2010) that allowed to damp the noise effects, reaching also a good balance between time and frequency resolution (evolfft function, RSEIS package, R Development Core Team).

About the SOM learning process and the map topological organization there are three principal changes. The first one concerns the choice of the triangle width parameter to calculate the weighted cross-correlation value between a couple of patterns. Since the entire SOM process is based on the cross-correlation (the discriminant function) calculated for each couple of code vectors (that identify a single neuron onto the map) and one input vector (an H/V spectral ratio), the choice of the triangle width value is capital to calculate the agreement between patterns including peak shifts. Triangle width has to be chosen considering the spectral frequency resolution. A narrow triangle allows to identify narrow peaks of the H/V function and generally it increases the number of clus-
ters on the map, while on the contrary a wide triangle produces a less “specialized” map and generally a few and more meaningful large clusters. A reasonable choice of the triangle width could be a number of frequency bins corresponding to a range of $0.5 – 1 \text{ Hz}$.

During the SOM training process, a wise choice of the frequency range allows to compare patterns (code vectors and input vectors) only in the range of interest, optimizing the organization of the map and then obtaining a more meaningful data clusterization. Usually the frequency range of interest goes approximately from $1$ to $10 \text{ Hz}$, but it could be extended or even reduced to about a hundred of frequency bins.

The third improvement in the field of SOM learning process concerns the topological properties of the map. A flat map has the disadvantage that the neurons along its edges do not have the same number of neighbours as the other neurons. This implies a non-homogeneous training process for each neuron on the map. The solution is to remove the edges of the map: a toroidal map is obtained connecting the upper and lower edges of the flat map, getting a cylinder, then the bases of the cylinder are merged. Toroidal maps avoid the data clusterization along the edges of the map and then allow a reliable organization of patterns onto the map.

The last improvement conceived the field of results interpretation. Once the training process is complete and the data (H/V spectral ratios) have been arranged onto the map, the analysis of clusters content starts. Generally there is a quite large cluster and an amount of small clusters that contain “outliers” H/V spectral ratios. Among these “outliers” there are the H/V spectral ratios affected by a considerable amount of noise. Each cluster is generally made up of more than one neuron, then the first step is to identify clusters. An algorithm of cluster recognition could be used. This algorithm belongs to the family of the hierarchical clustering algorithms and uses the weighted cross-correlation to calculate the similarity value of the code vectors of each couple of neurons. At each step one neuron is merged to another one, forming a new cluster, or it joins an already existing cluster. At any threshold similarity value correspond a number of clusters detected onto the map, the more the threshold is low, the large is the number of (small) clusters that are identified.

SOM process allows an efficient automatic analysis of the structure of the input space. Identifying the main cluster, in most cases, means being able to obtain the most stable shape of the H/V function for the investigated site, characterized by a quite low standard deviation (in amplitude) as required by SESAME guidelines (SESAME, 2004). Moreover SOM analysis allows to preserve the time information, so, being data presented to the neural network temporally ordered, considerations about data clustering process can help to understand if each cluster properties are stable in time or not.

**Outline of the SOM algorithm.** A Self Organizing Map (SOM) is an artificial neural network (ANN) that performs unsupervised competitive learning. Artificial neurons are arranged on a low-dimensional grid (generally a 2D map; Ultsch, 2003) and each neuron is described by a $n$-dimensional code vector, where $n$ is the dimensionality of input data. Each input vector could be for example a H/V spectral ratio computed in a given time window. When an input vector is presented to the network it causes a localized region or “bubble” of activity. Position and nature of this region usually change with the input vector and during the learning process. The first step to obtain a trained SOM map, once the size and geometry of the map are fixed, is the initialization of code vectors. The competition process follows when, for each input vector, the neurons compute their respective value of a discriminant function. The single neuron with the lower (or largest) value of the discriminant function is declared the winner of the competition (also called best matching unit, BMU). The effect of the competition process is that the continuous input space $X$ of activation patterns is projected onto the discrete output space $M$. During the subsequent cooperation process the BMU determines the spatial location of a topological neighborhood of excited neurons on the map. The neighbourhood function determines how strongly the neurons are connected to each other, it must be unimodal with the lateral distance between the BMU neuron and each other neuron (a typical choice is the gaussian function). The neighbourhood function’s value also depends on discrete time that iden-
tifies the iteration number; at every step the whole data set will be processed by the network examining each input vector in random order. Another parameter, called neighbourhood radius, defines the effective width of the neighbourhood function, or the degree in which the winning neuron affects its neighbourhoods. During the learning process the neighbourhood radius must be reduced monotonically with the regression step. The last step is called adaptive process; it enables to modify the code vectors of excited neurons to increase their individual values of the discriminant function in relation to the input vector. The adjustment takes place in a way that reinforces the answer of the BMU neuron with the application of similar input patterns, that allow the map training process and the data clustering.

Acknowledgements. This work was supported by PRIN 2007 project no. 2007PTRC4C_002 "Validazione di tecniche semplificate per la stima della amplificazione sismica di sito".

References
SESAME; 2004: Guidelines for the implementation of the H/V spectral ratio technique on ambient vibrations measurements, processing and interpretation. European research project.