Automatic Seismic Facies Classification with Kohonen Self Organizing Maps - a Tutorial

Atish Roy[1], Marcilio Matos[2], Kurt J. Marfurt[1]

[1]*Geology and Geophysics, University of Oklahoma, Norman, OK, United States*

[2]*Institute Militar de Engenharia, Brazil*

Abstract

The most popular seismic attributes fall into broad categories - those that are sensitive to lateral changes in waveform and structure, such as coherence and curvature, and those that are sensitive to lithology and fluid properties - such as AVO and impedance inversion. Unfortunately neither of these two attribute families works well in differentiating depositional packages characterized by subtle changes in stratigraphic column or the lateral changes in texture. Automatic seismic facies analysis aims to classify similar seismic traces based on amplitude, phase, frequency and other seismic attributes. This research shows how this technique helps to visualize the variation in rock properties and the possible relationship between these rock types and their seismic expression.

There are many different supervised and unsupervised clustering algorithms. In supervised training, the clusters are pre-defined and patterns are assigned to it in subsequent trainings. Unsupervised clustering is data driven without any a priori information. The self-organizing map (SOM) (Kohonen, 2001) is one of the most effective unsupervised pattern recognition techniques, and is generally used for the automatic identification of seismic facies.

We describe a SOM implementation that begins by over-defining the number of initial clusters from the input dataset. These "prototype vectors" are organized using Kohonen SOM neighborhood training rule. These clusters are projected onto a 2D or 3D latent space and are displayed against the continuous 2D or 3D colorbar. Then these clusters are compared back to the input seismic traces and colors are assigned to these seismic traces. The colored output map of seismic facies is then defined by the interpreter's eye and brain, in conjunction with well and other geophysical attributes.

Introduction

The final goal of seismic exploration is to make a reservoir model to understand the specific reservoir properties and map the heterogeneities present in the reservoir. The final hydrocarbon production depends on the porosity, permeability, fluid saturation and the extent and thickness of the reservoir. While the well data are sparse, the 3D seismic data cover a large area. The goal of the seismic interpretation is thus to interpolate and extrapolate the well data using the seismic data.

Since the seismic response depends on the reflectivity of the underlying rocks, change in seismic amplitude generally reflects change in the depositional environment, lithology and fluid present in the rocks. The description and interpretation of extracted seismic reflection parameters which include geometry, continuity, amplitude, frequency, and interval velocity, is known as seismic facies analysis (Mitchum, 1977). Within a certain interval, the seismic facies variation gives rise to changes in amplitude, phase and frequency of the seismic reflection. The objective of automatic clustering algorithm is to use such seismic as input and extract the natural clustering of the underlying facies.

Coleou et al. (2003) and their colleagues were among the first to publish SOM-based seismic waveform classification using a 1D latent space. More recently, Matos et al. (2009) showed the advantage of extending the latent space to 2D and 3D (with corresponding 2D and 3D colorbars) in delineating channels of a turbidities reservoir from the Campos Basin, offshore Brazil.

Several commercial products use the SOM algorithm to perform seismic facies analysis. However in this current paper we have approached a different coloring technique of the clusters with 2D using a Hue, Saturation, Value (HSV) Colorbar, which facilitate the visual identification of the seismic facies. We begin our tutorial with a review of the Kohonen Self Organizing map and the theory of cluster analysis. We demonstrate the mechanics of the algorithm using a simple 3D synthetic. Finally, we apply this technique to a tabular salt minibasin tectono-stratigraphic province (Diegel et al., 1995) of the northern Gulf of Mexico margin.

Kohonen Self Organizing Map

Kohonen (1982) presented his clustering and dimensionality reduction methods as a "self-organizing process" whereby a "simple network of adaptive physical elements" is made to resonate in a particular way with externally provided signals (a "primary event space") and tried to link these ideas with the functionality of the human brains (Murtagh, 1995). This popular neural network method was initially used in biology and computer science for data mining purposes; much later it has been used for an automated pattern-recognition technique in seismic exploration.

SOM (Kohonen, 2001) clusters data such that the statistical relationship between multidimensional data is converted into a much lower dimensional latent space that preserves the geometrical relationship among the data points. Mathematically each SOM unit within the latent space preserves the metric relationships and topologies of the multidimensional input data. SOM Prototype Vectors (PVs) or neurons have the same dimension as the input data (e.g. 12 dimensions for 12 amplitude values, or four dimensions for four texture attributes). These PVs are arranged in a regular low-dimensional grid or map (2D or 3D for our work), thereby topologically connecting a PV to its neighbors. By construction, SOM preserves the original topological structure within this dimensional attribute space, making it amenable for seismic facies analysis.

SOM Clustering Analysis

SOM closely relates to vector quantization methods (Haykin, 1999). Initially we assume that the input are represented by *J* vectors in a *N* dimensional vector space $\mathcal{R}^n$, $x_j=[x_{j1}, x_{j2}, x_{j3} .... x_{jN}]$ where *N* is the number of input attributes (or amplitude samples for "waveform" classification) and $j=1,2,...,J$ is the number of vectors analyzed. The objective of the algorithm is to organize the dataset of input seismic attributes, into a geometric structure called the SOM. SOM consists of neurons or Prototype Vectors (PVs) organized by a lower dimension grid, usually 2D, which are representative of the input data and have the same dimension of the input seismic attributes. PVs are also termed as SOM units and typically arranged in 2D hexagonal or rectangular structure maps that preserve the neighborhood relationship among the PVs, i.e., PVs close to each other are associated with neighbors input data. The number of these PVs in the 2D map determines the effectiveness and generalization of the algorithm.Considering 2D SOM represented by *P* prototype vectors m_i, $m_i= [m_{i1,} m_{i2} .... m_{iN}]$, where $i=1,2,...,P$ and *N* is the dimension of these vectors also being equal to the input vectors. The 2D SOM can be understood as a 2D sheet with the interconnected PVs imbedded on it. After assigning the PVs either in the hexagonal (Figure 2b) or in rectangular grid, we train these PVs (i.e., changing their individual weight values) in accordance to the input data. After SOM neighborhood training these PVs become clusters representing different classes of the input dataset.To illustrate this complex algorithm let us imagine a fanciful example. Figure 1(a) shows some fruits that have different properties. Let us consider two of their properties: their aspect ratio (shape) and peak frequency (color) and group these fruits according to these properties. After training, the fruits are arranged in three major groups; round and red fruits, elongated and green, round and blue, as shown in Figure 1(b). Now if we draw an analogy with a 2D SOM grid map, PVs represent the fruits; the shape and the color of the fruits are the attributes of the PVs and these different attributes help to arrange them into different groups of clusters.

Figure 2a shows 300 samples belonging to three distinct Gaussian distributions with the same standard deviation but different means. These 300 inputs vectors have three attributes or dimensions (*N*=3). Each sample vector in 3D space is cross-correlated with itself and all other vectors, resulting in a 3 by 3 covariance matrix. The first two Eigen values of this covariance matrix define the 2D latent space, which we shape with an 11x7 regular hexagonal (Figure 2b). Thus the total PVs or the number of over-defined classes are 77 (P=77). Each of these individual PVs is denoted by m_i and is of dimensionality 3 (same as the input vector). The weight of the PVs associated with each lattice element *i* lying inside the neighborhood radius σ (t) of the 2D map is updated with successive trainings.

Along SOM training process, an input vector is initialized and is compared with all *N* dimensional PVs on the 2D grid latent space. The prototype vector with the best match (Winner PV) will be updated as a part of SOM

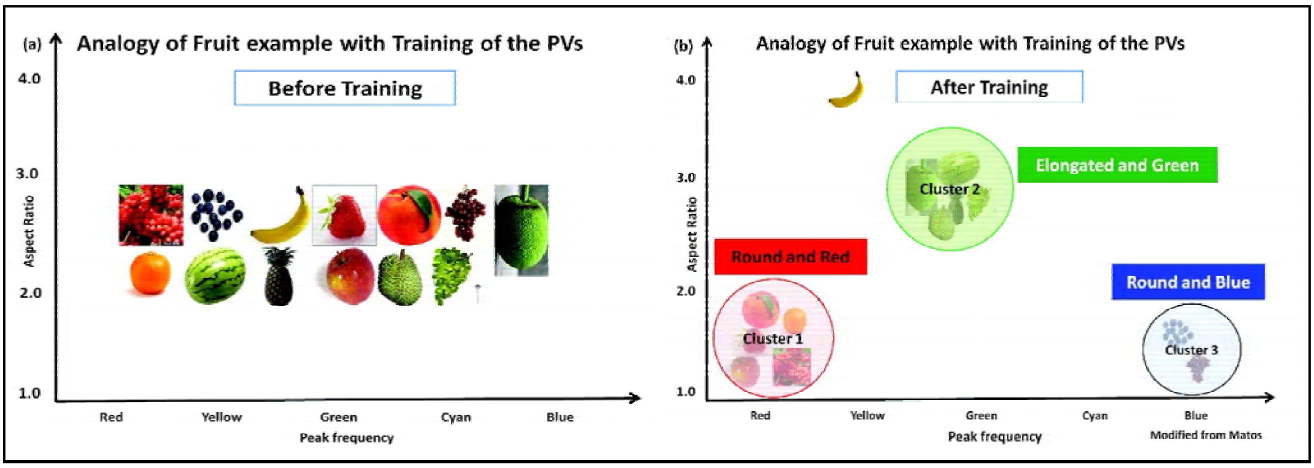

Fig.1 Example of different fruits illustrating unsupervised clustering analysis. (a) The unorganized fruits before training; (b) Clustering of the fruits after training considering their two properties (color and aspect ratio).

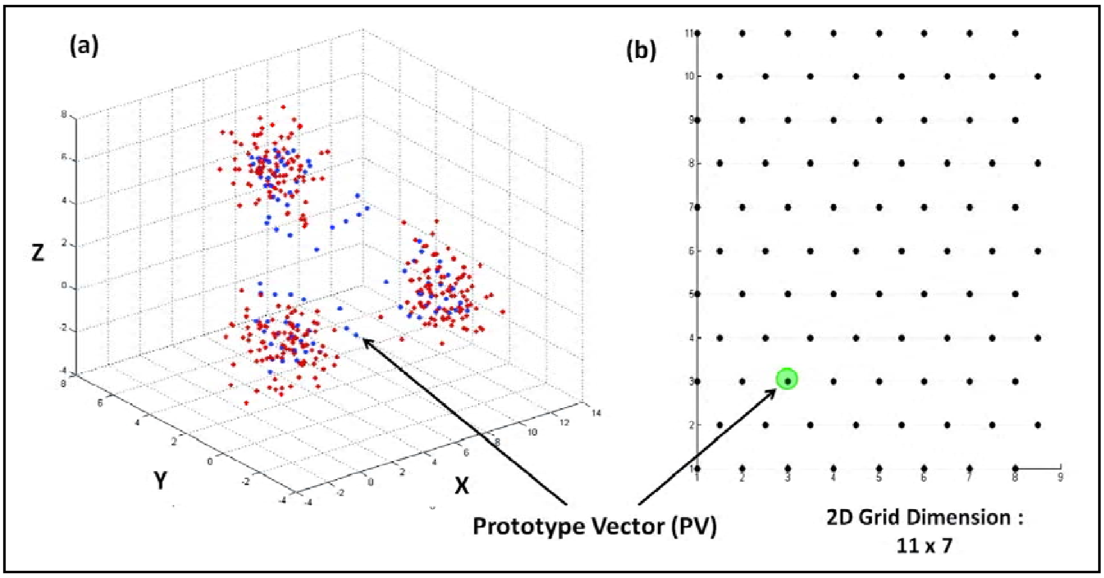

Fig. 2(a) A plot of samples corresponding to three normal Gaussian distributions with the same standard deviation but different means. The red plus signs represents the original data points and the blue dots the Prototype Vectors in 3D; (b) the initial unorganized PVs plotted on hexagonal grid.

neighborhood training. Small neighborhood of PVs around the "winner" prototype vector are also updated (Figure 3a). It means the PVs lying within a distance of "σ(t)" is excited and updated. With successive iteration, this neighborhood radius decreases. Thus, the weight of the "winner" prototype vector is brought closer to the input trace with the training.

With this background, the SOM training algorithm can be sequentially written in the following steps (Kohonen, 2001):

Step 1: Consider an input vector, which is randomly or systematically chosen from the set of input vectors.

Step 2: Compute the Euclidean distance between this vector and all PVs, $i=1, 2,\ldots p$. Then the prototype vector, which has the minimum distance to the input vector, is termed as the "Winner" or the Best Matching Unit, (equation 1).

$$\| x-m_b \| = MIN \{ \| x-m_i \| \} \qquad \ldots(1)$$

Step 3: Update the "winner" prototype vector and its neighbors. The updating rule for the weight of the i th PV inside the neighborhood radius σ(t) is given by equation 2(a) and outside σ(t) given by equation 2(b).

$$m_i(t+1) = m_i(t) + \alpha(t)\, h_{bi}(t)[x - m_i(t)] \quad if \| r_i - r_b \| <= \sigma(t) \ldots(2a)$$

$$= m_i(t) \quad if \| r_i - r_b \| >= \sigma(t) \ldots(2b)$$

where the neighborhood radius defined as σ(t) is pre defined for a problem and decreases with each iteration t; r_b and r_i are the

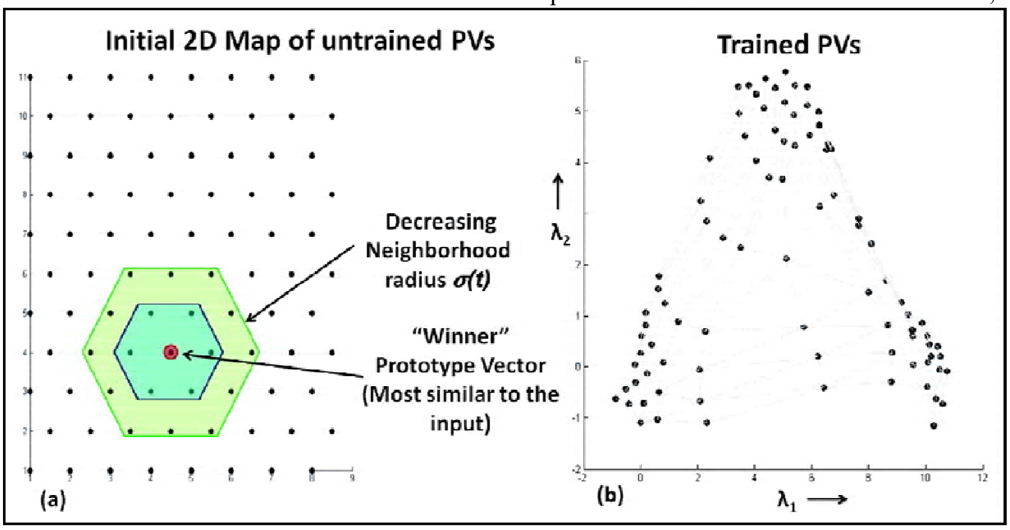

Fig. 3 2D map of the PVs (a) before and (b) after the training. Around the "Winner PV" the larger green hexagon represents a neighborhood at an early iteration while the smaller blue hexagon represents a neighborhood at a later iteration. The results can be interpreted as forming three clusters, thus classifying the input data properly.

position vectors of the winner PV m_b and the i th PV m_i respectively. We also define $h_{bi}(t)$ as the neighborhood function, $\alpha(t)$ as the exponential learning function and T as the length of training. $h_{bi}(t)$ and $\alpha(t)$ decreases with each iteration in the learning process and they are represented by the following equations 3 and 4,

$$h_{bi}(t) = e^{-(\|r_b - r_i\|^2/2\sigma^2(t))} \quad \text{... (3)}$$

$$\alpha(t) = \alpha_0 \left(\frac{0.005}{\alpha_0}\right) t/T \quad \text{... (4)}$$

The training of the PVs for the above dataset is illustrated in Figure 3. The green hexagon represents a neighborhood, $\sigma(t)$ at an early iteration, while the smaller blue hexagon represents a neighborhood at a later stage. Although 77 prototype vectors are used, after training they group themselves into three separate clusters in the 2D latent space.

Step 4: Iterate through each learning steps (Step 1-3) until the convergence criterion (which depends on the pre defined lowest neighborhood radius) is reached.

Step 5: Project the trained PVs onto 2D latent space using Principal Component Analysis, and color-code by using 2D or 3D gradational colors (Matos et al. 2009). For 2D latent space, we will use an HSV model with Hue, $\mathcal{H}$

$$\mathcal{H} = \tan^{-1}\left(\frac{\nu - \frac{1}{2}}{\mu - \frac{1}{2}}\right)$$

and Saturation, $\mathcal{S}$ is defined as

$$\mathcal{S} = [(\mu - \tfrac{1}{2})^2 + (\nu - \tfrac{1}{2})^2]^{\frac{1}{2}}$$

where μ and ν are the projected components and the first two unit length Eigen vectors V^1 and V^2 The input data contain three separate classes. Thus figure 4a also shows three separate clusters in this latent space. These new set of PVs are colored by the 2D HSV colorscale with equations 5 and 6 (Figure 4b). This coloring scheme will be used for seismic facies maps as discussed in the remainder of the paper.

Application of SOM on a synthetic data

To verify the effectiveness of the above workflow for waveform classification, we apply it on the 3D synthetic model shown in Figure 5. The 3D volume is composed of four different waveforms shown in Figure 5c. The location of the different waveforms A, B, C and D are highlighted both on the map and crossline view (Figure 5a and b) of the synthetic model. The data are sampled at 2ms, consist of 50 inlines by 26 crosslines and are 30ms (16 samples) long.

From the input 1300 traces, we mathematically define a 26 x 7 grid corresponding to 182 PVs (Figure 6a), each 16 samples long. The magnitude of each prototype vector after SOM training eventually is modified to one of the classes present in the original data. After training, these prototype vectors on the 2D maps get the shapes of the different waveforms present in the input dataset. These trained waveforms are then plotted according to their original location (Figure 6b).

Examining the Unified Matrix – essentially a distance matrix formed from the trained PVs (Figure 6c) we note four distinct blue areas separated from each other corresponding to four different waveforms or seismic facies.

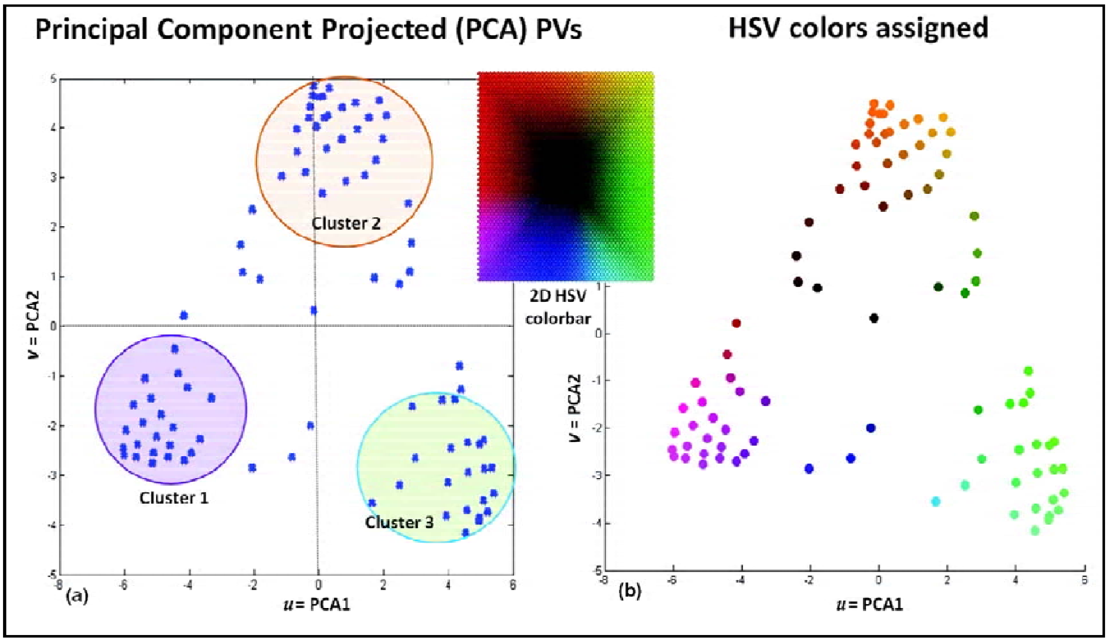

Fig. 4 Interpreter - based clustering by mapping a 2D latent space against a 2D colorbar. The trained PVs are projected in a different (a) 2D latent space (here by Principal Component Analysis) and then (b) colored by the 2D HSV colorscale using the equation for Hue and Saturation (Eqs: 5 and 6)

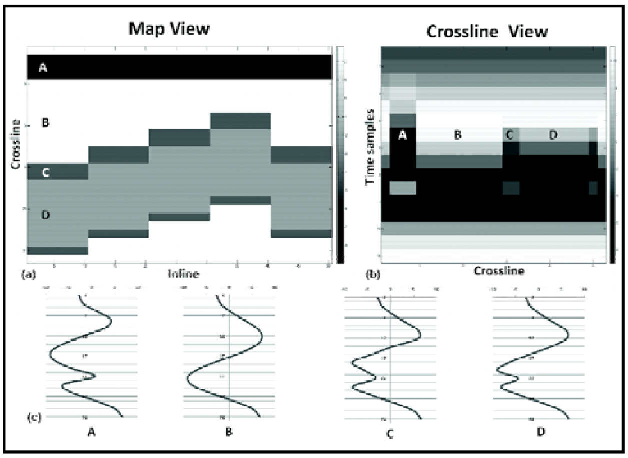

Fig.5 The synthetic data with a sample rate of 2ms and consisting of 50 inlines by 26 crosslines and are 30ms (16 samples) long (a) Map view and (b) Crossline view. (c) The different waveforms A, B, C and D plotted for their whole length; these four waveforms are highlighted on both the map and crossline view 6(a) and (b).

Project the trained PVs onto a latent space, defined by Principal Component Analysis and color-coding using the HSV 2D gradational colorbar with equations 5 and 6 (Figure 7). We see that the different waveforms fall into four distinct clusters.

The resulting facies map shows four different colors corresponding to the four different waveforms (Figure 8a) in the input dataset. The waveforms are superposed on the 2D map (Figure 8b) to show how the different colors are assigned to the different waveforms. The waveforms with their respective colors are explicitly shown in Figure 8c.

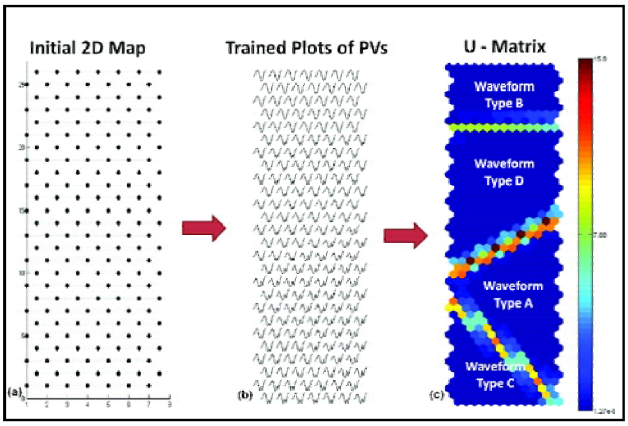

Fig.6(a) 2D map of the untrained PVs. (b) Line plots of the trained PVs on 2D and re arranging them in order. (c) The Unified Matrix (U-matrix) - a distance matrix formed from the trained PVs, shows four distinct blue areas separated from each other corresponding to the four different waveforms.

Application of SOM on a dataset from Northern Gulf of Mexico

Our study area is within the tabular salt mini-basin, tectono-stratigraphic province (Diegel et al., 1995) of the

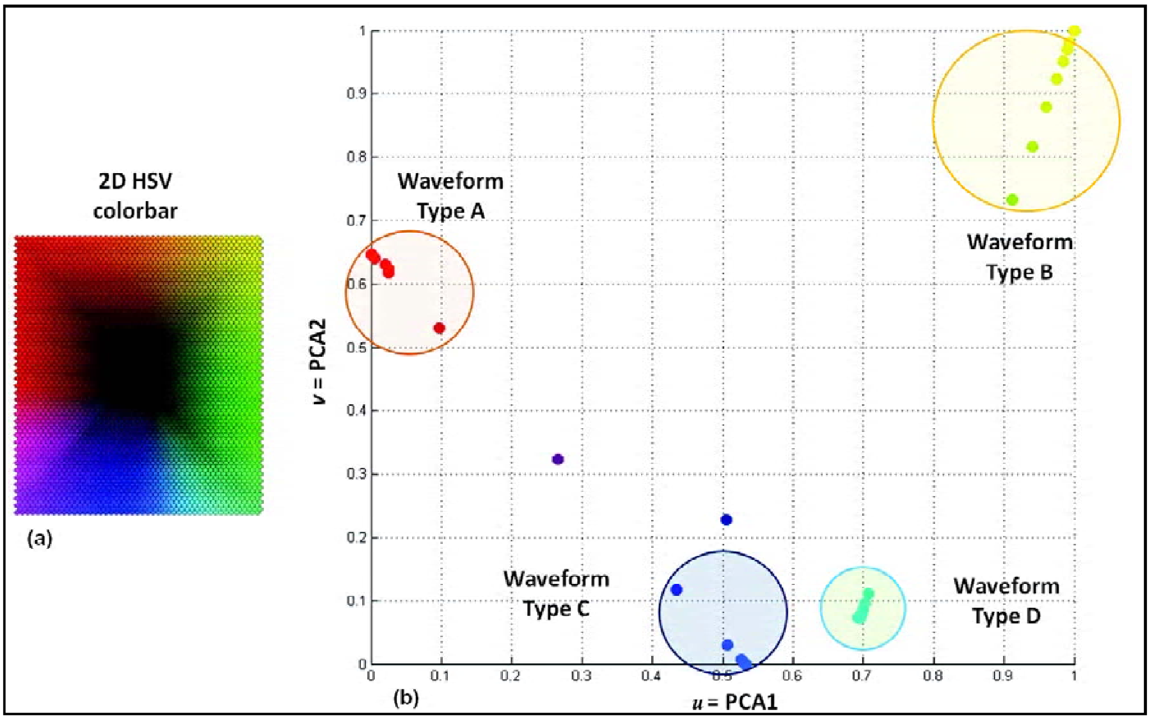

Fig. 7 Interpreter-based clustering by mapping a 2D latent space against a 2D colorbar using PCA. The different input waveforms A, B, C and D are thus colored differently. These colored PVs will be compared back to the input data and different waveforms will be classified separately.

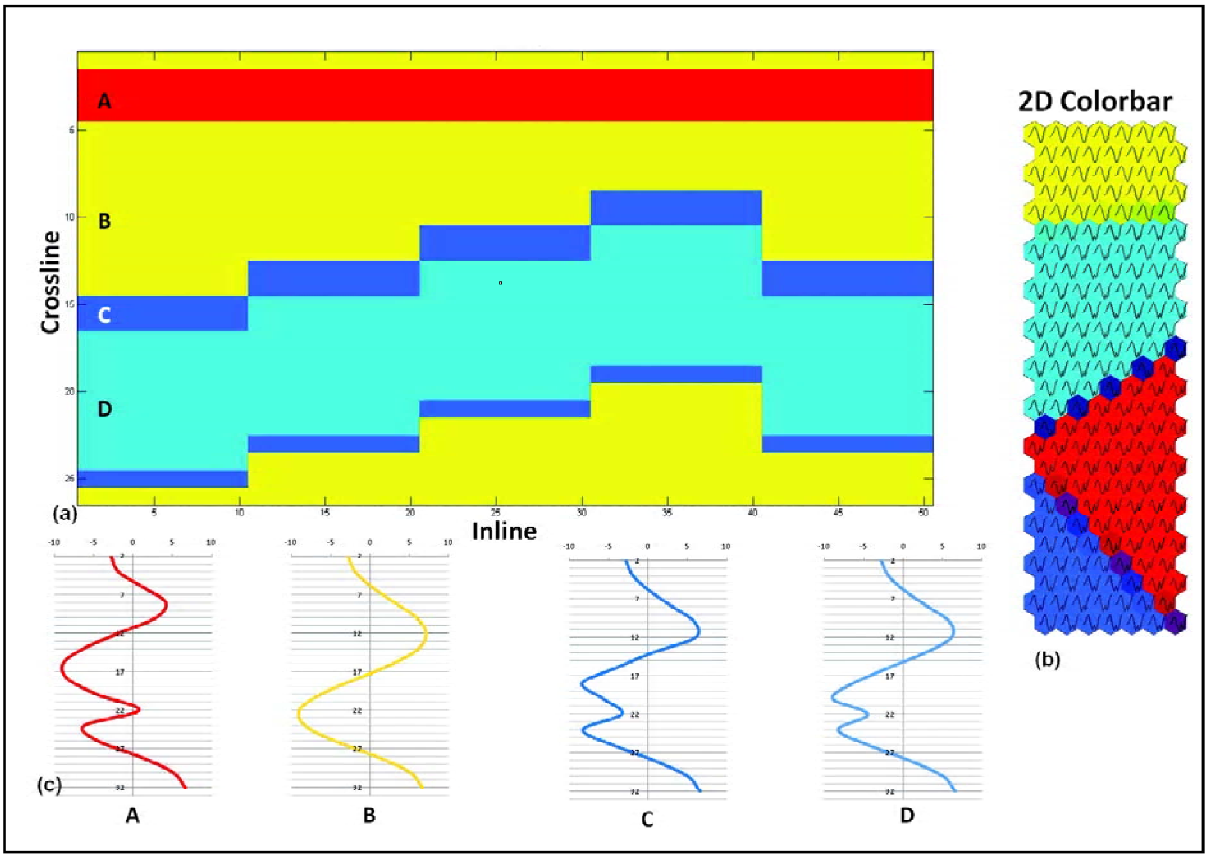

Fig. 8(a) The seismic facies map shows four different colors corresponding to the four different waveforms that constitute the synthetic dataset. (b) The waveforms are superposed on the 2D colormap to show how the different colors are assigned to the different waveforms. (c) These four waveforms are explicitly shown with corresponding colors assigned to it.

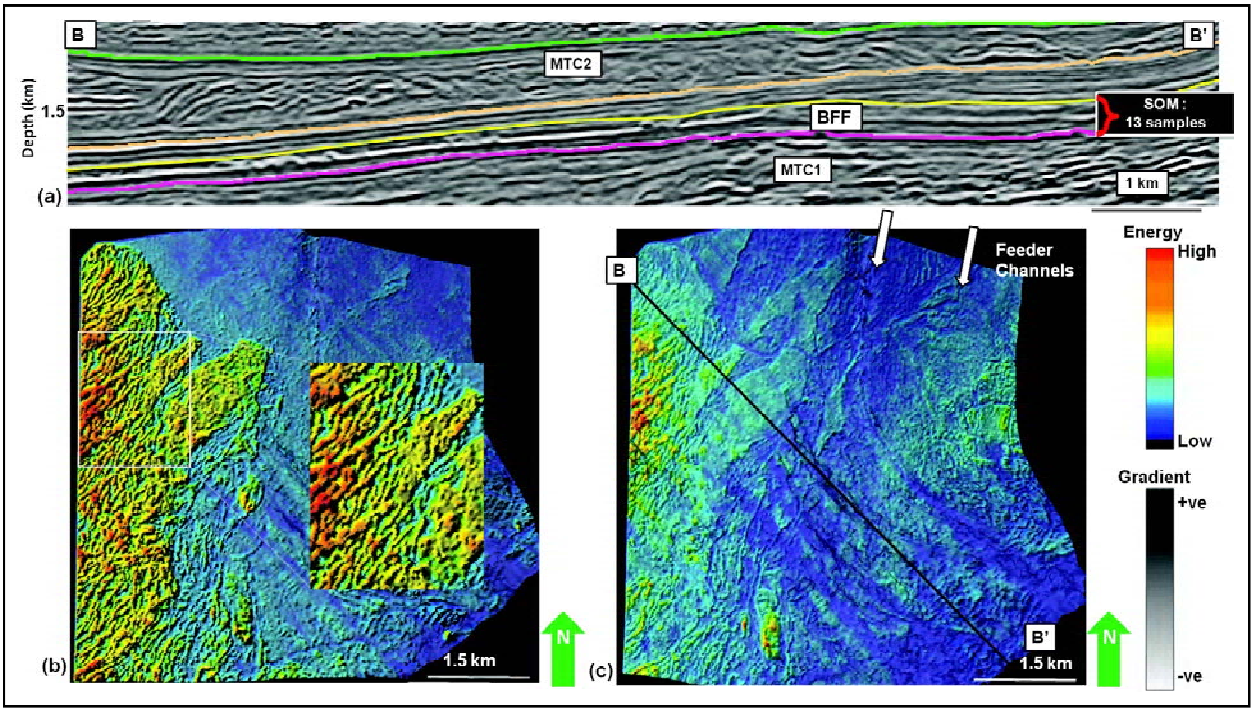

Fig. 9(a) Seismic section BB' showing the basin Floor fan (BFF) zone (b) stratal slices generated by co-rendering coherent energy and inline amplitude gradient volume highlighting the basin floor fans and (c) the feeder channels (Sarkar et al., 2010)

Northern Gulf of Mexico (GoM) margin. There are primarily four depositional environments for deep water systems: channels, thin bed-levees, basin floor fans or sheets and mass transport complexes (MTCs). The sediments in the salt minibasins follow in a "fill and spill" depositional process (e.g., Winker, 1996) and this process preserves the basin floor fan and channel-levee sediments.

A 100 km^2 prestack depth migrated 3D seismic data volume covers almost this entire salt minibasin. Sarkar et al., (2010) describes a suite of different MTCs preserving the basin floor fan in the system. They co-render strata, slices of coherent energy and inline amplitude gradient to highlight the basin floor fans (Figure 9b) and their feeder channels (Figure 9c). For our analysis we have taken 13 sample input data above a flattened horizon above the MTC1 (Figure 9a) in the basin floor fan zone. In the vertical seismic section BB' the basin floor fan lies just below the horizon marked in yellow. Our objective is to construct a seismic facies map to cluster the different waveforms in the analysis zone.

Figure 10a represents the trained set of PVs which are projected onto a 2D latent space constructed using PCA projection. The HSV color is already applied to the trained PVs. Figure 10b highlights a small part of trained PVs. Each seismic trace from the input data is compared with all the trained PVs using the same rule,

$$\| x-m_b' \| = min \{ \| x-m_i' \| \} \quad ...(7)$$

where m_i' is the trained i th PV and x is the input data vector and m_b is the nearest PV to the input vector x. We find to which PV the input seismic trace is most similar to (i.e. having the minimum Euclidean distance). The input seismic data is then assigned the color of m_b'. Since trace 1 in Figure 10c is most similar to the blue colored PV the location of trace 1 is assigned blue and similarly traces 2 and 3 are assigned green and red, respectively. Repeating the process for the input seismic traces, we obtain the seismic facies map as shown in Figure 11.

Figure 11a shows the seismic facies map after SOM analysis of the basin floor fan deposits. This map represents the 13 sample input dataset. Similar colors represent similar seismic characteristics, with the 2D HSV colormap (Figure 11b) showing how different waveforms are assigned different colors. Note the magenta waveforms are totally different when compared to the green waveforms. The SOM thus colorcodes different waveforms according to their similarities. The yellow interpretations in Figure 11a are drawn to help better understanding this map. From the works of Sarkar et al.,(2010) we know that the feeder channel cuts through the shale and form the fan deposits. Note that the feeder channels are shown in green compared to the general blue (pelagic shale) background. Generally coarser deposits form nearer and finer deposits further away from the source, which can be seen with the variation of the color (from brown to pink) in the basin floor fan as we move across from east to west. Figure 11c is the overlay of the seismic facies map on the flattened

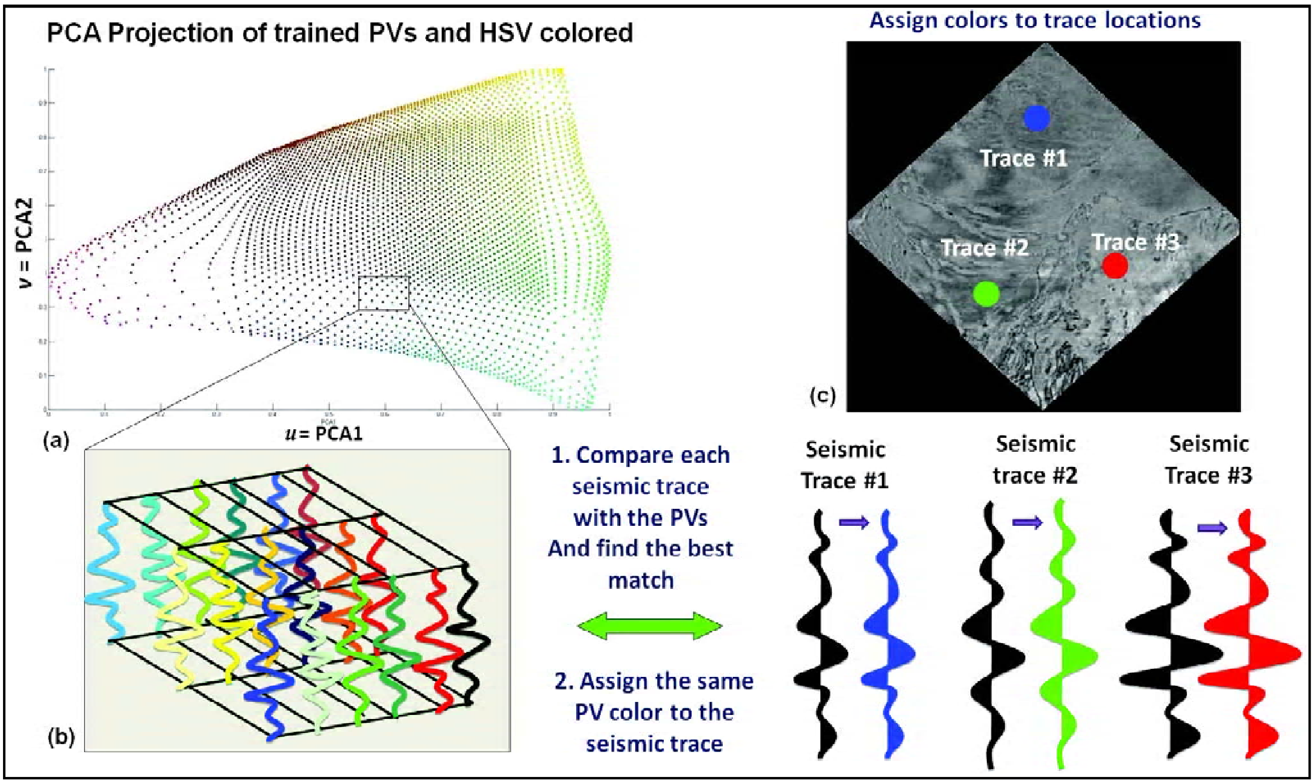

Fig. 10 Schematic representation of the generating a seismic facies maps from a set of (a) trained and colored PVs. (b) Highlighting a small set of colored PVs. (c) The seismic traces 1, 2 and 3 get the color of that PV to which it matches the best. In a similar fashion we come up with a colormap where different color represents different seismic waveforms representing different depositional facies.

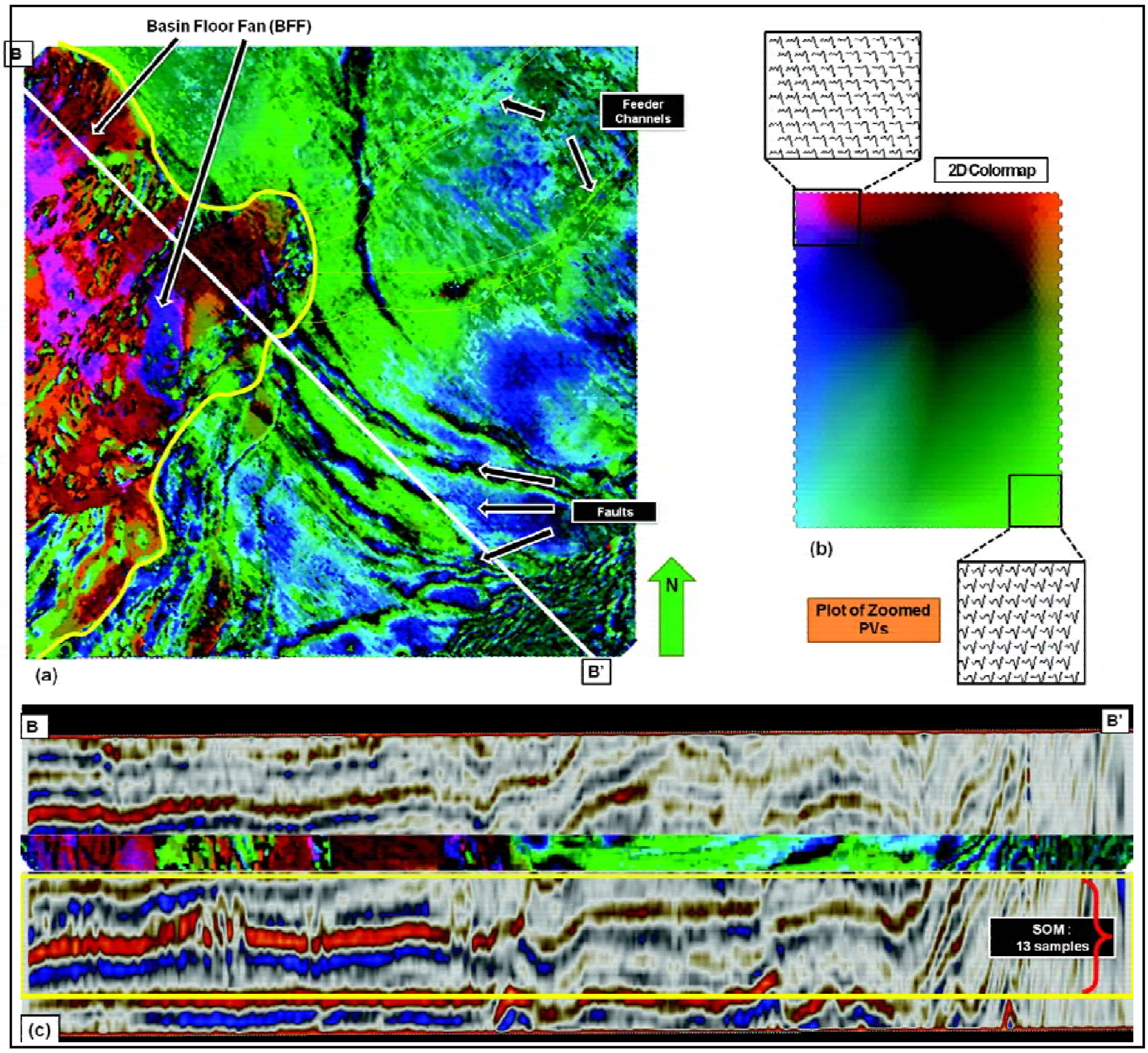

Fig. 11 SOM clustering analysis. (a) Seismic facies map from SOM training considering 13 sample input seismic amplitude dataset. The yellow interpretation in marked to help in visualizing the basin floor fan (BFF) and their feeder channels better. There is a variation in the color within the BFF which can be inferred due to the change in depositional nature within the fan. The pelagic shale deposits are represented by the general blue background. (b) The 2D HSV colorbar shows different colors are assigned to the different waveforms. The waveforms to which the green color is assigned is different from the waveforms with pink colors. (c) The seismic facies map at BB' being overlaid on the vertical seismic section to see how different colors are assigned to different waveforms.

seismic section showing how different colors correspond to the variation in the seismic waveforms. Thus the general geologic nature of the subsurface can be understood from this seismic facies map.

Discussion

If the seismic data are processed correctly, waveform changes are associated with changes in stratigraphy. Thus by "waveform classification" we can have a decent knowledge of the subsurface geology and the reservoir heterogeneities. This algorithm is also helpful when using multi-attribute data as input to SOM. For that we can take different correlating attributes for a required reservoir zone and then run SOM on it to see the results (Roy et al., 2010). The waveforms will be classified with the constraints being the given correlating attributes.

However the results should be validated with the logs from wells and they should be tied more accurately to the rock physics properties. Then we will be able to understand the physical reasons for the variation in colors in the seismic facies map. A limitation to this 2D color mapping is the folding of the PVs when plotted in the 2D latent space, because of which multiple PVs can have the same colors. In that case different classes may not be separated into different colors.

Supervision into the SOM algorithm from the well logs should result in a more confident "waveform classification". From this case study and previous studies it can be inferred that the SOM gives a decent knowledge of the subsurface depositional environment and can be used while exploring new areas where good geological information is not present.

References

Coleou, T., M Poupon, and K. Azbel, 2003, Unsupervised seismic facies classification: a review and comparison of techniques and implementation: The Leading Edge, 22, 942-953.

Cvetkovic, M., N. Pralica, S. Falconer, and K. Marfurt, 2008, Comparison of some algorithms for acquisition footprint suppression and their effect on attribute analysis: SEG Expanded Abstract.

Diegel, F. A., J. F. Karlo, D. C. Schuster, R. C. Shoup, and P.R. Tauves, 1995, Cenozoic structural evolution and tectono-stratigraphic framework of the northern Gulf coast continental margin: AAPG Memoir 65, 109-151.

Duda, O, Richard., Hart, E, Peter., and Stork, G, David., Pattern Classification. Second Edition, Wiley- Interscience.

Kohonen,T., 2001, Self Organizing Maps,3rd ed.: Springer-Verlag.

Marroquin, D, I., Brault, J, Jean., and Hart, Bruce., 2009. A visual data mining methodology for seismic facies analysis: Part 1 – Testing and comparison with other unsupervised clustering methods. Geophysics vol. 74

Marroquin, D, I., Brault, J, Jean., and Hart, Bruce., 2009. A visual data mining methodology for seismic facies analysis: Part 1 – Application to seismic data. Geophysics vol. 74

Matos, M. C., K. J. Marfurt., and P. R. S. Johann, 2009, Seismic color Self-Organizing Maps: 11th International Congress of the Brazilian Geophysical Society, Expanded Abstracts.

Murtagh .F., Hernandez M., 1995., The Kohonen Self-Organizing Map Method : An Assessment.

Peyton, L., R. Bottjer, and G. Partyka, 1998, Interpretation of incised valleys using new 3-D seismic techniques: using spectral decomposition and coherency: The Leading Edge, 17, 1294-1298.

Roy, A., Matos, M. C., and K. J. Marfurt., 2010, Applying Self Organizing Maps of Multiple Attributes, an example from the Red-Fork, Anadarko Basin: SEG Expanded Abstract.

Sarkar, S., K. J. Marfurt., and Slatt, R., 2010, Generation of sea-level curves from depositional pattern as seen through seismic attributes-seismic geomorphology analysis of an MTC-rich shallow sediment column, northern Gulf of Mexico, The Leading Edge, 29, 1084-1091.

Suarez, Y., K.J. Marfurt, and M. Falk, 2008, Seismic attribute-assisted interpretation of channel geometries and infill lithology: a case study of Anadarko Basin Red Fork channels: SEG Expanded Abstract.

Wallet, C. Brad., Matos, M. C., Kwiatkowaski., Suarez Yoscel., 2009, Latent space modeling of seismic data : An overview. The Leading Edge, Vol. 28, No. 12, pp. 1454-1459.

Winker, C. D., 1996, High-resolution seismic Stratigraphy of late Pleistocene submarine fan ponded by salt-withdrawal mini-basins on the Gulf of Mexico continental slope: Proceedings of the 28th Offshore Technology Conference, 619-628.