Seismic characterization of the Middle Jurassic Hugin sandstone reservoir in the southern Norwegian North Sea with unsupervised machine learning applications for facies classification

Satinder Chopra[1*], Thang Ha[2], Kurt, J. Marfurt[2] and Ritesh Kumar Sharma[1].

Summary

Because they allow us to integrate the information content contained in multiple seismic attribute volumes, machine learning techniques hold significant promise in the identification and delineation of heterogeneous 3D seismic facies. However, considerable care must be taken in choosing not only the appropriate, but also in their scaling. Sometimes such exercises are carried out mechanically, resulting in compromised interpretations and discouraging results. We examine some of the more well-established unsupervised machine learning techniques such as principal component analysis (PCA) and k_{means} clustering, as well as some less common clustering techniques like independent component analysis (ICA), self-organizing mapping (SOM), and generative topographic mapping (GTM) as applied to a seismic data volume from the southern Norwegian North Sea. We find that the machine learning methods can provide increased vertical and spatial resolution. However, machine learning is also good at enhancing noise and artifacts. For this reason, the interpreter needs to ensure the data are adequately conditioned, the assumptions on which some of the techniques being applied are based are met, and finally, the most appropriate technique among those discussed in this paper is utilized.

Introduction

Located in block 15/9 in the southern Norwegian North Sea on the continental shelf, in 80 m of water, the Volve oilfield is situated approximately 200 km west of Stavanger (Figure 1a). It was discovered in 1993 when exploration well 15/9-19SR was drilled and found oil in the Middle Jurassic Hugin sandstone formation. The objective for 15/9-19SR was to test the hydrocarbon potential of the Paleocene age Heimdal Formation which forms the main reservoir in the adjoining Sleipner Øst area. The well showed the Heimdal to be dry but encountered oil in the Hugin sandstone. Initially estimated oil and gas reserves stood at 78.6 million barrels of oil and 1.5 billion cubic metres of gas, but with the drilling of observation well 15/9-F11-A, the oil reserves were increased by 8.8 to 9.4 million barrels. The production from the oil field started in February 2008 and reached a plateau of 56,000 barrels per day of 29.1 °API oil. The field delivered a total of 63 million barrels of oil spread over 8.5 years and reaching a recovery rate of about 54%. It was decommissioned in September of 2016.

The stratigraphy for the reservoir interval is shown in Table 1.

Figure 1b shows a segment of a seismic section with three markers overlaid. The Base of Cretaceous Unconformity (BCU) represents the separation of the syn-rift depositional sequence from the post-rift depositional sequence and covers large areas in the North Sea. It is easily identified on surface seismic data and well logs and is an important marker. The dashed black marker represents the base of the Hugin sandstone reservoir, which forms a combined structural and stratigraphic trap, with depths varying between 2750 to 3120 m. These sandstones are not preserved over the entire survey. The western side of the structure is heavily faulted and the communication across the faults is uncertain.

Our goal is to determine the facies distribution within the Hugin sandstone reservoir and develop a better understanding of

System	Group	Formation	Lithology
Jurassic	Viking	Draupne Fm.	Claystone, minor Limestone
		Heather Fm.	Claystone
	Vestland	Hugin Fm.	Sandstone, minor Claystone and Limestone (**Reservoir**)
		Sleipner Fm.	Sandstone-Claystone intercalation. Minor Coal

Table 1 Stratigraphic sequence for the reservoir level of the Volve field area. (Modified from Sen and Shankar, 2019)

[1]SamiGeo, Calgary | [2]The University of Oklahoma, Norman
* Corresponding author, E-mail: Satinder.Chopra@samigeo.com
DOI: 10.3997/1365-2397.fb2021089

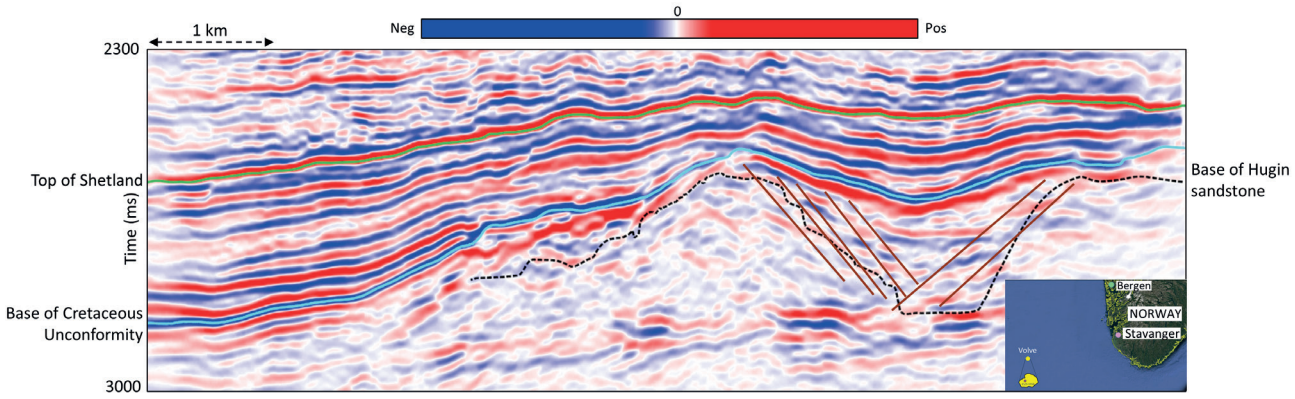

Figure 1 Segment of a section from the seismic volume with three markers overlaid. The 'Base of Hugin sandstone' shown with the black dashed line represents the base of the Middle Jurassic sandstones, which form the reservoir in the Volve Field. These sandstones are not present over the full survey. The inset shows an index map of the location of Volve oilfield in the southern Norwegian North Sea, generated using Google Earth. The approximate shape of the field is shown in the inset (not drawn to scale).

the techniques that may be employed for doing so. Because the porosity of sandstone influences the impedance, the ideal workflow is to conduct prestack simultaneous impedance inversion of the seismic data to obtain P-impedance and V_P/V_S ratio. Even with such estimates, it is useful to augment such 'quantitative' attributes based on the well-established correlation of the seismic amplitude response to changes in reflectivity, with 'softer' measures provided by complex trace analysis, spectral decomposition, and texture attributes that are sensitive to not only the vertical but also the lateral variation in the reflectivity. Furthermore, for reasons of insufficient S-wave logs, or simply for cost and time constraints, in many cases prestack inversion is not performed. Therefore, instead of the absolute P-impedance from prestack simultaneous impedance inversion, we generated relative acoustic impedance and used that in the attribute mix instead. Thus, for the delineation of the Middle Jurassic Hugin sandstone facies, the eight attribute volumes used for this exercise were relative acoustic impedance, instantaneous envelope, sweetness, GLCM (grey-level co-occurrence matrices) energy, peak magnitude, and spectral magnitudes at 35 Hz, 40 Hz and 45 Hz.

1. *Relative acoustic impedance* is computed by continuous integration of the original seismic trace with the subsequent application of low-cut filter. The impedance transformation of seismic amplitudes enables the transition from reflection interface to interval properties of the data, without the requirement of a low-frequency model.
2. *Instantaneous envelope* is a measure of the instantaneous energy of the analytic seismic trace, independent of phase, and provides information on intensity of reflections.
3. *Sweetness* is a 'meta-attribute' or one computed from others, which in this case is the ratio of the envelope to the square root of the instantaneous frequency. A clean sand embedded in a shale will exhibit high envelope and lower instantaneous frequency, and thus higher sweetness, than the surrounding shale-on-shale reflections.
4. *GLCM* or grey-level co-occurrence matrix energy is a measure of textural uniformity in the data. If the reflectivity along a horizon is nearly constant, it will exhibit high GLCM energy.
5. *Peak magnitude* represents the spectral magnitude of the seismic data at the peak frequency. Similarly, the magnitude at the peak frequency, from which the average magnitude of the entire spectrum of the data is subtracted, is called 'peak magnitude above average', and can also be generated. It emphasizes the anomalous (typically tuning) response of the data and thus is helpful for at-a-glance interpretation.
6. *Spectral magnitude* is the magnitude of each spectral component ranging within the seismic bandwidth of the data at a specific frequency increment.

The choice of the attributes used in unsupervised learning relies on their data distribution as well as any correlation that might exist between them. The selection of the input attributes can be done in different ways including geologic insight, quantitative correlation with geologic features of interest, and previous successes obtained for similar objectives in other basins. In this example, we will use our geologic insight to choose the candidate attributes. Details of how this selection was made is not the focus of this paper and will be discussed in another publication.

Seismic facies classification using machine learning techniques

Machine learning uses mathematical operations to learn from the similarities and differences in the provided data and make predictions. Besides the supervised and deep learning machine learning techniques (Figure 2), there are two broad families of unsupervised machine learning algorithms. The first algorithm family includes dimensionality reduction algorithms such as PCA and ICA. When plotted against a 2D colour bar, the interpreter may be able to identify clusters, but the algorithm output is a continuum of data in a lower dimensional space. The second, unsupervised classification algorithm family attempts to explicitly cluster the data into a finite number of groups that in some metric 'best represent' the data provided. k_{means} clustering is one such process. Before the analysis, there is no interpretation assigned to any given group; rather, 'the data speak for themselves'. However, the choice of input attributes biases the clustering to features of interpretation interest. Biasing the training data to favour geologic features of interest (e.g., by providing a disproportionate number of voxels exhibiting a bright-spot anomaly) also provides interpreter control of the output. We also show the application of *self-organizing mapping* (SOM) and *generative topographic mapping* (GTM) to the Volve data volume.

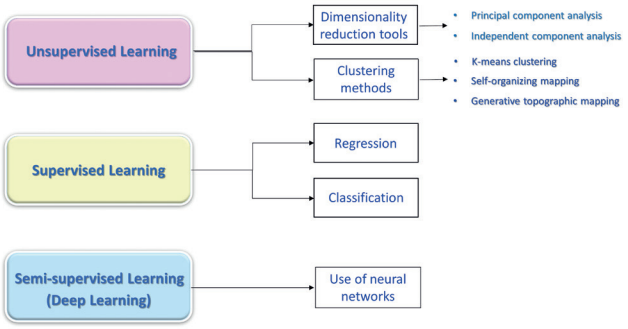

Figure 2 Classification of machine learning techniques.

The objective of this exercise is to enhance our understanding about the application of unsupervised machine learning (ML) techniques for facies classification including principal component analysis (PCA), independent component analysis (ICA), k_{means} clustering, self organized mapping (SOM), and generative topographic mapping (GTM). Our goal is to understand the sensitivity to data conditioning and its impact on the eventual crossplotting analysis, the assumptions for the application of some of the techniques, and what we need to do should such assumptions not be satisfied. We provide a visual comparative performance of the ML techniques for each of these issues.

Principal component analysis

Principal component analysis (PCA) is perhaps the most commonly used dimensionality reduction tool. If the input data do not exhibit a Gaussian distribution, the resulting principal components will also be skewed, thus reducing the 'resolution' of the colour near the peak of the distribution during corendering or crossplotting (Ha et al., 2021).

Many of our attributes are coupled through the underlying geology, such that a fault may give rise to lateral changes in waveform, dip, peak frequency, and amplitude. Such 'redundant' images provide increased confidence in the existence and extent of a given geologic feature. Less desirably, many of our attributes are coupled mathematically, such as alternative measures of coherence (Barnes, 2007) or of a suite of closely spaced spectral components. The amount of attribute redundancy is measured by the covariance matrix. The element C_{mn} of an N-by-N covariance matrix is then simply the crosscorrelation between the m^{th} and n^{th} scaled attribute over the volume of interest containing R length-N data vectors. By convention, the first step is to order the eigenvalues from the largest to the smallest. The first eigenvector is a linear combination of scaled attributes that best represents the variance in N data volumes. Commonly, the jth eigenvalue λ_j is normalized by the sum of all the eigenvalues such that $0 \le \overline{\lambda}_j \le 1$. If we wish to represent the N attribute volumes by a smaller subset of linear attribute combinations, we choose a cut-off value, say a fraction ε of the variation expressed by the first eigenvalue, and ignore all eigenvectors and principal components j where

$$\lambda_j < \varepsilon\lambda_1, \tag{1a}$$

or alternatively

$$\overline{\lambda}_j < \varepsilon \sum_{q=1}^{q=j-1} \overline{\lambda}_q . \tag{1b}$$

To analyse N attribute volumes, each voxel is represented by a length-N attribute vector. The projection of (crosscorrelation) the length-N attribute vector against the length-N first eigenvector, $\mathbf{v}_1$

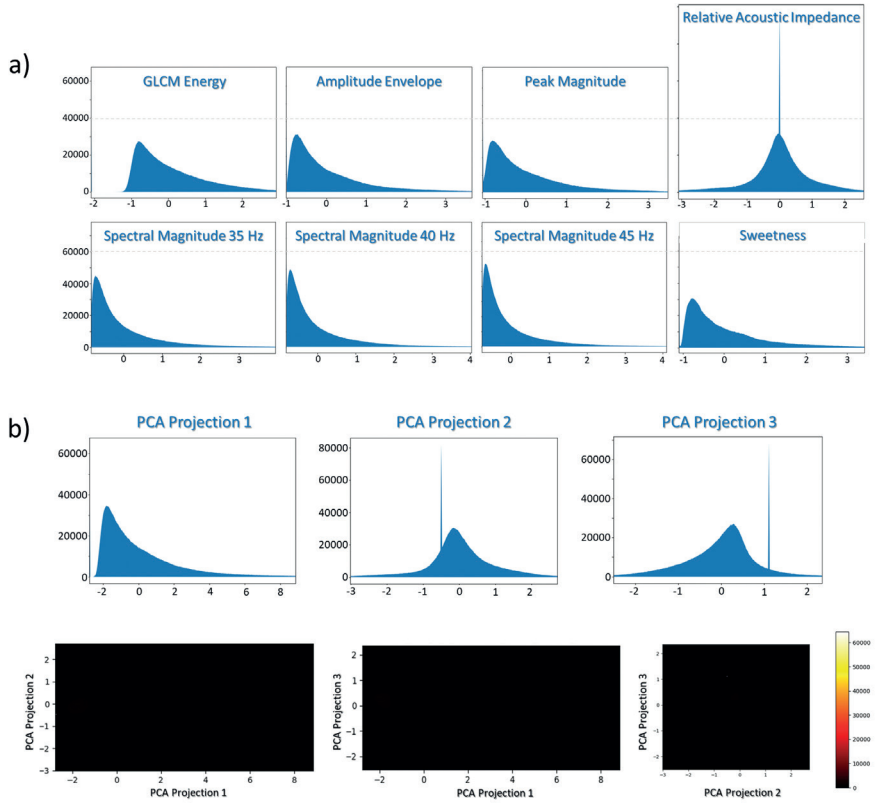

Figure 3 (a) Attributes selected for use in machine learning processes and their amplitude distributions. (b) Crossplots of PCA projections 1, 2, and 3 against one another computed after z-score transformation. The spikes seen in the data prevent the visibility of the cluster points on the crossplots.

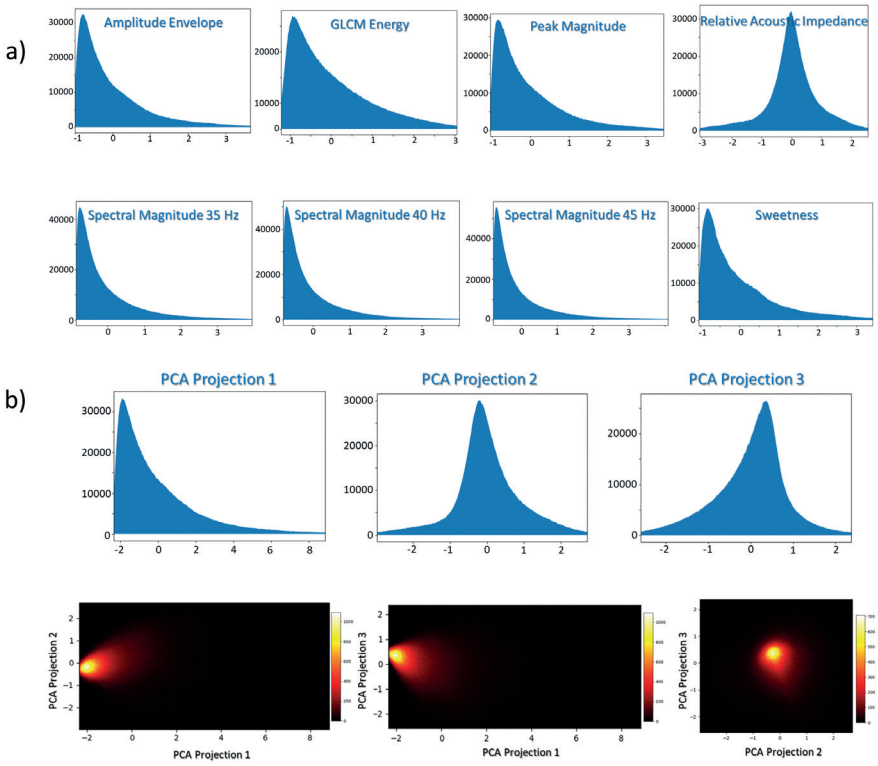

Figure 4 Analysis of datasets shown in Figure 1 after conditioning (despiking, etc.) (b) Crossplots of PCA Projection 1, 2, and 3 against one another computed after z-score transformation. After removal of spikes in the data the cluster points on the crossplots become visible.

results in a crosscorrelation coefficient called the first principal component of the data set (PC1), which captures the largest amount of data variance. Because the first eigenvector v_1 best represents the variance of the data analysed (e.g., between two picked horizons) as a whole, PC1 will map out the major features in the data. Eigenvector $\mathbf{v}_2$ is paired with the second-highest eigenvalue, $\overline{\lambda}_2$. When crosscorrelated with the data vectors, it provides the second principal component. PC2. $\mathbf{v}_1$ and $\mathbf{v}_2$ define a plane in *N*-dimensional attribute space that best represents (in a least-squares sense) the *N* attribute volumes. Similarly, the third eigenvector $\mathbf{v}_3$ will be perpendicular to the plane defined by $\mathbf{v}_1$ and $\mathbf{v}_2$. In our experimentation we have found the first three principal components can represent the vast majority of the data (~ 75 – 90%). The display of PC4 and beyond are increasingly noisy and provide less geological information. Besides, we can only effectively corender up to three components on a display. Therefore, our analysis will be based on the first three principal components.

In Figure 3a we show the amplitude distributions of the eight attributes used in this exercise. Interestingly, barring the relative acoustic impedance, all the attributes show skewed distributions. The amplitude distribution for relative acoustic impedance exhibits a spike. Such spikes can originate from the dead or muted zones. Sometimes, they are also found to originate in the inversion process and are seen in the computed porosity attribute as clipped values corresponding to zero or 100%. It is critical to exclude dead traces and mute zones from the analysis, and while many attributes have the same mute zones as the original data, some attributes such as relative acoustic impedance may need to increase the mute zones to minimize edge effects. In Figure 3b we show the crossplots for PCA projection 1, 2, and 3 against one another, where to our dismay we notice that they are blank. On closer examination we realize that this is due to a spike in one of the attributes that prevents the cluster points from showing up on the crossplot.

To address this issue, we removed all data vectors containing a value close to the (user-defined) spike value, and then regenerated the displays shown in Figure 3 as Figure 4. The spike has gone away completely, and the amplitude level has come down (Figure 4a). Also, the amplitude levels that were dwarfed compared with the spike amplitude have now been enhanced. More importantly, the crossplots now exhibit cluster points as expected (Figure 4b).

Figure 5 shows stratal slices 64 ms above the base of the Hugin sandstone from the RGB (red, green, and blue) corendered principal components (PC1, PC2 and PC3), generated before and after preconditioning for removal of the histogram spikes. This level has been chosen as it represents the optimum surface representing Hugin sandstone. Note the increased colour range and how constant colour patches in Figure 5a are seen with greater lateral detail in Figure 5b, assuming that maximum variance along the stratal slice is geologically meaningful. A similar observation can be made for the equivalent RGB corendered displays (Figure 6) made for the independent components to be discussed in the next section. In Figure 7 we display equivalent stratal displays from the input seismic data volume as well as the attributes used in the computation of the principal component analysis. This has been done to help the interpreters gauge how the outcomes of the various algorithms differ from and add information to the raw data.

Normalization of seismic attributes

The first step in multi-attribute analysis is to subtract the mean of each attribute from the corresponding attribute volume. If

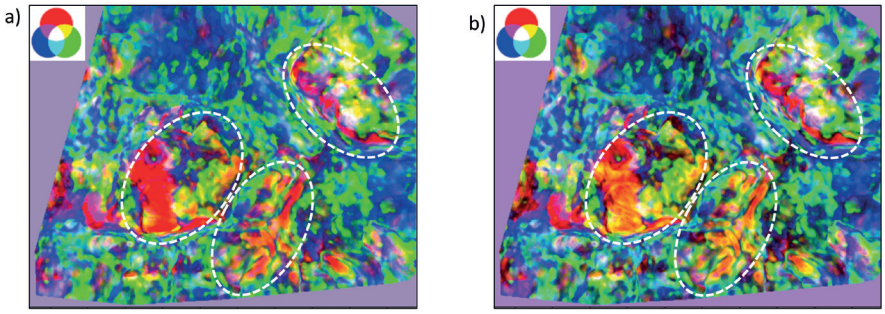

Figure 5 Stratal slice from RGB corendered principal components (PC1, PC2 and PC3) computed using z-score normalization of the input seismic attributes, (a) without, and (b) with preconditioning. Notice the display shows more colour variation (and hence more details) after preconditioning.

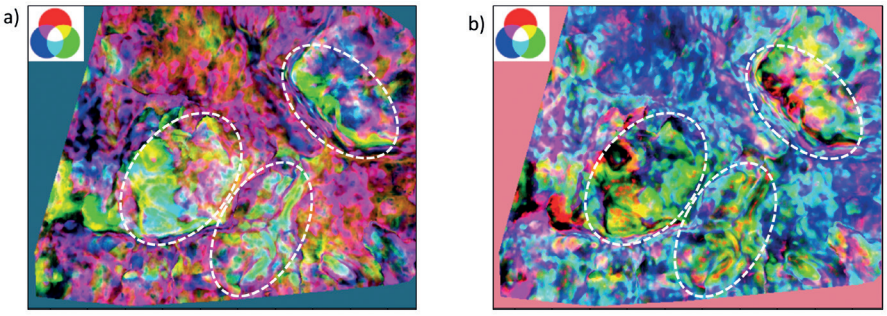

Figure 6 Stratal slice from an RGB corendered independent components (IC1, IC2 and IC3) computed using z-score normalization of the input seismic attributes, (a) without, and (b) with preconditioning. Notice the display shows more colour variation after preconditioning.

the attributes have radically different units of measure, such as frequency measured in Hz, envelope measured in mV, and coherence without dimension, a z-score normalization is required, where the difference of the attribute and its mean is divided by the standard deviation. While the z-score transformation is a widely used normalization scheme, it assumes that the input data exhibit a distribution close to a normal or Gaussian distribution. In actual practice, many of the seismic attributes exhibit skewed, peaked, or flattened distributions (Figure 3a). The simple z-score normalization does not compensate for skewness or kurtosis of the input distribution.

To address this issue, we scale each length-N attribute vectors **x** to obtain scaled attribute vectors **y** using

$$y_{nr} = b_n(x_{nr} + a_n) \tag{2}$$

where for attribute n a z-score normalization $a_n = -\mu_n$ is called the shift factor, and $b = \frac{1}{\sigma}$ is the scale factor. For Gaussian data μ_n is the mean, and σ_n is the standard deviation. The z-score transformation has only two parameters, the mean μ_n and the standard deviation σ_n and is a linear transformation.

For non-Gaussian distributions we adopt the algorithm described by Ha et al. (2021), who demonstrated the value of (when appropriate) a logarithmic transformation that can help to overcome the shortcomings of a z-score transformation and recast equation (2) into

$$y_{nr} = c_n\left\{\ln\left[b_n(x_{nr} + a_n)\right]\right\} \tag{3}$$

where c_n is a scale factor in the logarithmic domain. Such a transformation is equivalent to first linearly transforming the input, followed by the application of a logarithmic function. Even though this transformation has three parameters (a_n, b_n, and c_n) instead of two in the z-score transformation (a_n and b_n), some constraints need to be defined so that the logarithmic transformation can reshape the input data distribution close to a Gaussian distribution.

Ha et al. (2021) reshape the data distribution curve in terms of three anchor points: the value x_p at the p^{th} (e.g., 2.5%) percentile, the value $x_{1\text{-}p}$ at the $(1\text{-}p)^{th}$ (e.g., 97.5%) percentile, and the mode, x_{mode}, representing the peak of the distribution. The objective of this reshaping process is that after the logarithmic transformation, the transformed left and right anchor points are symmetric about zero, with the peak of the transformed distribution located at zero. When this process was implemented, they found that a slightly skewed distribution was obtained instead of a symmetric one. In reshaping, the logarithmic transformation moves the peak of the transformed distribution away from the peak of the original distribution. To avoid this, Ha et al. (2021) adopt an iterative procedure for computation of transformation parameters by recomputing the peak anchor point of the original distribution, x_{mode} from the peak of the transformed distribution at every iteration. For ensuring convergence of the process to the peak, rather than hovering around it or get close to it, the average peak of all iterations is computed, after which a linear scale factor is also computed to shift the mean of the reshaped distribution to zero.

We illustrate the adoption of the logarithmic transformation in our analysis of the input attributes in Figure 8, where equivalent plots are displayed to those shown in Figure 4. Notice the distributions more closely approximate to a Gaussian for each of the input attributes (Figure 8a), and the greater spread of the cluster points on the principal component projection crossplots (Figure 8b).

As in Figures 5 and 6, we show the stratal slices, extracted from RGB corendered principal components using the traditional z-score normalization (Figure 9a) and the logarithmic normalization (Figure 9b). Greater colour variation is seen on the display

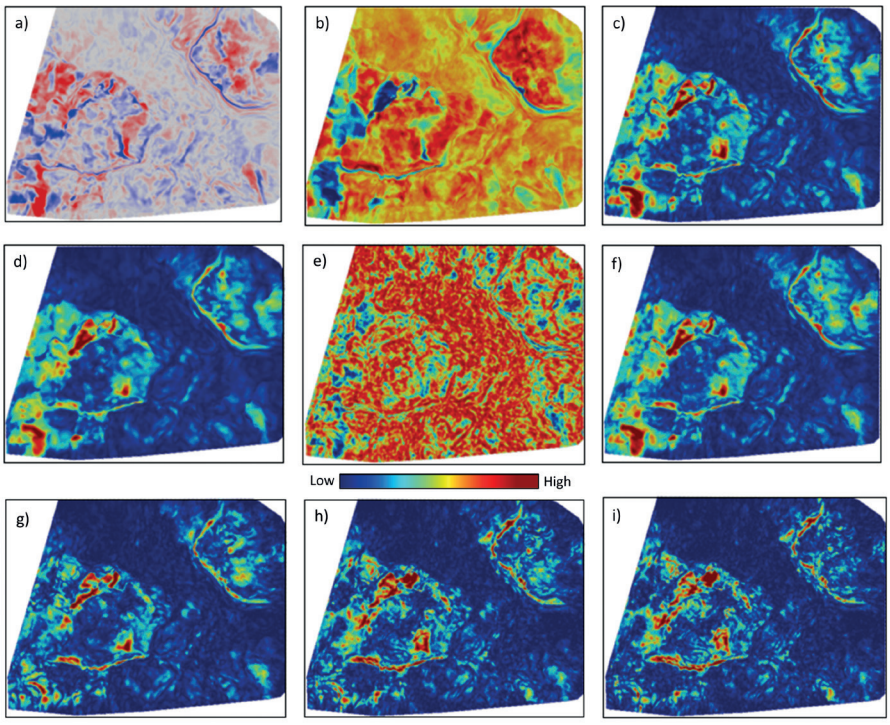

Figure 7 Stratal slices 64 ms above the Base of Hugin sandstone marker extracted from the (a) seismic, (b) relative acoustic impedance, (c) amplitude envelope, (d) sweetness, (e) GLCM energy, (f) peak magnitude, and spectral magnitude at (g) 35 Hz, (h) 40 Hz, and (i) 45 Hz volumes.

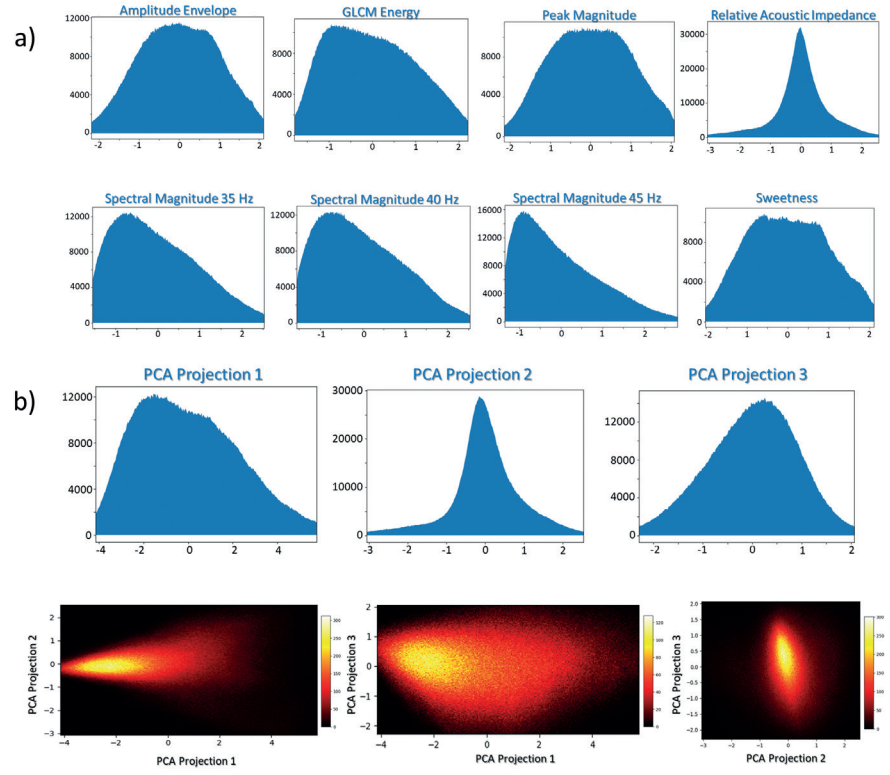

Figure 8 Analysis of datasets after conditioning (despiking, etc.) and logarithmic transformation. (b) Crossplots of PCA Projection 1, 2, and 3 against one another computed after logarithmic transformation. They better utilize the full range of 2D or 3D colour table, thereby delineating smaller geologic details.

in Figure 9b, which we interpret as exhibiting more lateral variability and is assumed as geologically meaningful.

When we non-linearly transform the data through the logarithmic transformation, the data are stretched near the peak and squeezed near the 'tail' of the distribution. If the objective is to define subtle features within a reservoir that exhibit a seismic expression similar to the neighbouring layers, we will obtain better colour resolution using the logarithm scaling. However, if the objective is to define features such as a salt diapir, karst collapse, or gas chimney, whose seismic expression is quite different from the surrounding facies, then we want to preserve those strong differences by using a simple z-score transformation.

Independent component analysis

Independent component analysis (ICA) is an elegant machine learning technique that separates multivariate data into independent components, without the requirement of a Gaussian

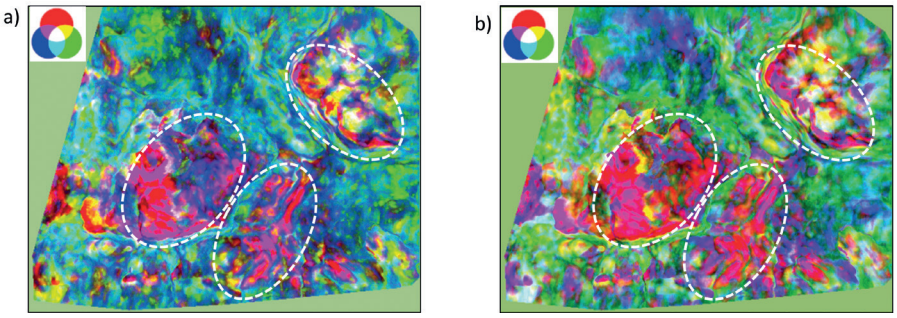

Figure 9 Stratal slice from an RGB corendered principal components (PC1, PC2 and PC3) computed using (a) z-score, and (b) logarithmic normalization of the input seismic attributes.

distribution for data going into the analysis. The other differences between ICA and PCA are that the independent components are not orthogonal, and their order is not defined, in that the first, second and third ICAs are ordered by visual examination, and are not mathematically ordered in the process as in PCA (Honorio et al., 2014; Lubo-Robles, 2018; Chopra et al., 2018).

Given a combination of different seismic attributes as input data, ICA attempts to find the 'mixer' that acts on several independent components, which is mathematically cast as a matrix equation, and solved using higher order statistics. Figure 10 illustrates the differences between the principal and independent component analysis. We demonstrate its application to the Volve multiattribute seismic data using both the z-score and logarithmic normalization (Figure 11), wherein the resultant independent components exhibit better resolution and separation of the geologic features with the latter process.

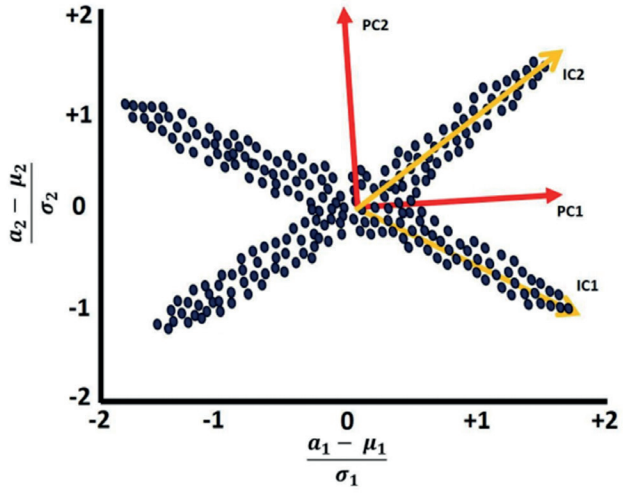

Figure 10 Differences between Principal Component Analysis (PCA) and Independent Component Analysis (ICA). The normalized attributes ***a*1** and ***a*2** (scaled by their means and standard deviations) are shown on the two axes. The first eigenvector **v1** is a line that least-squares fits the data cloud and best represents the variance of the data. PC1 is a projection of each data point onto **v1**. The second eigenvector **v2** is perpendicular to **v1** and for two dimensions these two eigenvectors best represent the data. In contrast, the independent components IC1 and IC2 are latent variables whose order is undefined, and they are not orthogonal to each other (Hyvärinen and Oja, 2000). To compute the independent components, each data point is projected onto the whitened eigenvectors **v1** and **v2**, and then projected onto the unmixing matrix W. (After Lubo-Robles, 2018)

k-means clustering

k-means clustering is one of the simplest clustering algorithms and is available in most seismic interpretation software. k_{means} organizes a given distribution of length-N attribute vectors at R voxels, $\mathbf{x}_r$, where $r = 1, 2, \ldots R$, into a desired number of k clusters. The clustering process begins by assigning at random k centroids which can serve as centres of the groups we wish to form, where each centroid defines one cluster. Next, the distance between each data point and the centroid of that cluster is calculated. A point may be within a cluster if it is closer to the centroid in that cluster than any other centroid. As some reorganization of the points in different clusters has taken place, the centroids are recalculated for each cluster. These two steps are carried out iteratively, until there is no more shifting of the centroids, and the process has converged. The calculation of distance between the centroid and the data points referred to above is the traditional Euclidean distance, which assumes there is no correlation between the classification variables. If there is no correlation, then the classification variables would exhibit a spherical shape of the clusters in crossplot space. In many cases, this is not found to be true, as the classification variables exhibit clusters that are elliptical in shape, and hence are correlated. In such cases, the traditional *k*-means clustering method might not achieve convergence and hence fail. To avoid this problem, we measure distance in the orthogonal principal component space rather than use Euclidean distance computed in the original non-orthogonal attribute space. We find that *k*-means clustering method using distances in the

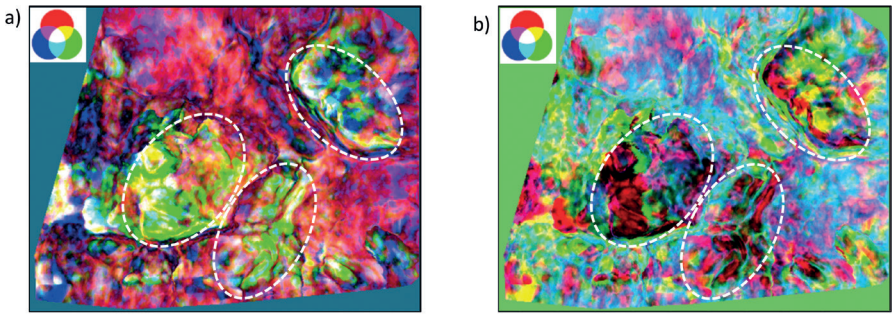

Figure 11 Stratal slice from an RGB corendered independent components (IC1, IC2 and IC3) computed using (a) z-score, and (b) logarithmic normalization of the input seismic attributes.

principal component space correctly classifies nonspherical and nonhomogeneous clusters.

Suppose eight input attributes are used for k_{means} clustering into six output clusters. When these six clusters are plotted against a simple rainbow colour bar, they may not show how the clusters are related to one another. Explicitly, does the red cluster lie next to the orange cluster or the blue cluster in N-dimensional attribute space? An alternative way is to project the cluster centres onto the three principal axes, yielding $k^{(1)}_{means}$, $k^{(2)}_{means}$, and $k^{(3)}_{means}$ as outputs, which can then be integrated/corendered using RGB. Doing this will enable clusters that are closer together to be displayed with similar colours, while clusters that are far apart will be displayed with contrasting colours. This strategy reduces the need to estimate the number of clusters, because redundant clusters will now have similar colours.

In Figure 12 we show a stratal slice comparison for the z-score and logarithmic normalization for k_{means} clustering using the new distance metric generated with eight clusters. We see a better distribution of coloured patches on the display that used logarithmic normalization (as seen in the highlighted areas of the display), which are a possible representation of the different facies in the data at that level.

Self-organizing maps

Like k_{means}, *self-organizing mapping* (SOM) is a technique that generates a seismic facies map from multiple seismic attributes, again in an unsupervised manner. In contrast to k_{means}, SOM defines its initial cluster centroids in an *N*-dimensional attribute data space and uses the first two eigenvectors of the covariance matrix to least-squares fit the data with a plane (Kohonen, 1982, 2001). Grid prototype vectors (also called neurons) defined in this plane, are attracted to data out of the plane, deforming it into a 2D surface called a manifold that better fits the data. After convergence, the *N*-dimensional data are projected onto this 2D surface, and are turn mapped against a 2D plane or 'latent' (hidden) space defined by axes SOM-1 and SOM-2, in which the interpreter either explicitly defines clusters by drawing polygons, or implicitly defines clusters by plotting the results against a 2D colour bar.

Figure 13 shows the equivalent stratal display to the ones shown earlier, extracted from the crossplot generated between the SOM-1 and SOM-2 volumes using a 2D colour bar shown alongside. Figure 13a shows the display for the SOM computation carried out when the input attributes were normalized using the z-score approach, and Figure 13b displays the equivalent display when the input attributes were normalized using the logarithmic transformation. The clusters seen on the display in Figure 13b are better defined in terms of more colour separation and distinct definition than the ones seen in Figure 13a as well as those shown earlier from PCA and ICA analysis or the k_{means} clustering display.

Generative topographic mapping

Although popular as an unsupervised clustering technique for its straightforward implementation and its low computation cost, the Kohonen self-organizing map has limitations. There is no theoretical basis for selecting the training radius, neighbourhood function and learning rate as these parameters are data dependent (Bishop et al., 1998; Roy, 2013). No cost function is defined that could be iteratively minimized and would indicate the convergence of the iterations during the training process, and finally

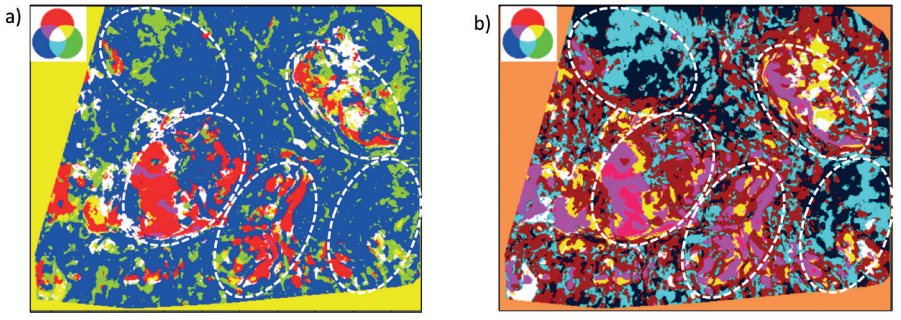

Figure 12 Stratal slice from an RGB corendered k_{means} clustering components (1, 2 and 3) computed using (a) z-score, and (b) logarithmic normalization of the input seismic attributes.

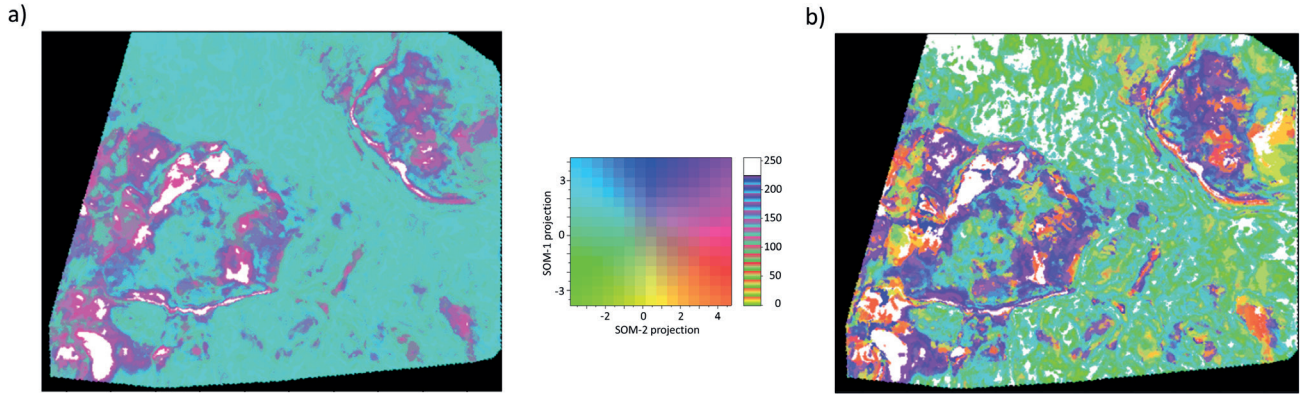

Figure 13 Equivalent stratal slices from SOM1 vs SOM2 crossplot volume, where SOM projections were computed using (a) z-score, (b) logarithmic normalization of the input seismic attributes. The white pixels are clipped data corresponding to outliers, which was necessary to enhancing the colour contrast of the greater part of the image. The clusters seen on the logarithmic normalization exhibit better spatial resolution than the equivalent display where z-score normalization was used.

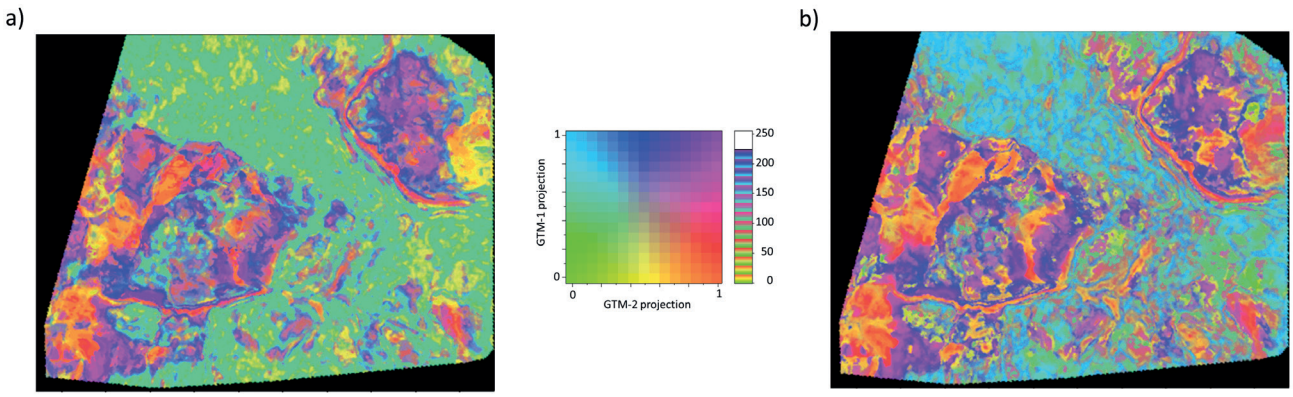

Figure 14 Equivalent stratal slices from GTM1 vs GTM2 crossplot volume, where GTM projections were computed using (a) z-score, (b) logarithmic normalization of the input seismic attributes. The clusters seen on the logarithmic normalization display exhibit better spatial resolution than the equivalent display where z-score normalization was used.

no probability density is defined that could yield a confidence measure in the final clustering results. Bishop et al. (1998) developed an alternative dimensionality reduction technique called a *generative topographic mapping* (GTM) algorithm that provides a probabilistic representation of the data vectors in latent space.

The GTM method begins with an initial array of grid points arranged on a lower dimensional latent space**,** e.g., the first two or three principal components or the ICA components. Each of the grid points are then nonlinearly mapped onto the lower dimensional non-Euclidean curved surface defined by K centroids $\mathbf{m}_k$ of the N-dimensional Gaussian functions with a fixed variance $1/\beta$ that best represent the R data vectors. At each iteration, the variance $1/\beta$ is decreased and the Gaussian centroids $\mathbf{m}_k$ moved until we reach convergence. Roy (2013) and Roy et al. (2014) describe the details of the method and demonstrate its application for mapping of seismic facies to the Veracruz Basin, Mexico.

As it may have become apparent from the descriptions above, the PCA, ICA, SOM and GTM techniques project data from a higher dimensional space (8D when eight attributes are used) to a lower dimensional space which may be a 2D plane or a 2D deformed surface. Once they are projected onto a lower dimensional space, the data can be clustered in that space, or interactively clustered with the use of polygons.

After the application of GTM to the data at hand, Figure 14 shows the equivalent stratal display to the ones shown earlier, extracted from the crossplot generated with GTM-1 and GTM-2 volumes using a 2D colour bar shown alongside. Figure 14a shows the display for the GTM computation carried out when the input attributes were normalized using the z-score approach, and Figure 14b displays the equivalent display when the input attributes were normalized using the logarithmic transformation. More spatial resolution in terms of colour is seen on the clusters.

Clusters seen on the display in Figure 14b are better defined in terms of more colour separation and distinct definition, than the ones seen in Figure 14a as well as those shown earlier from PCA and ICA analysis or the k_{means} clustering display.

Conclusions

We have found that editing and scaling can significantly impact the performance and colour resolution of multiattribute projection and classification techniques. Spikes in the data associated with mutes and no-permit zones need to be excluded from the analysis. We find that when our goal is to differentiate subtle features in the data that both limiting the zone of analysis to the geologic formation of interest, to limit the number of facies analysed, and logarithmic scaling of the data to stretch the attribute response provides the best delineation (colour resolution) of the geologic features. In contrast, if our goal is to delineate single seismic facies (e.g., salt diapirs exhibiting anomalously low coherence, low RMS amplitude, high GLCM entropy) that is represented by the tails of the data distribution, then we should apply a simple z-score scaling to preserve the tails.

Application of PCA, ICA, k_{means}, SOM and GTM techniques to the same data allowed us to assess their relative strengths. We found that principal component analysis provided more convincing results (greater differentiation of facies) than k_{means} clustering. ICA results look better than PCA results in terms of better colour detail and separation of the geologic features. Both GTM and SOM show promising results, with GTM having an edge over SOM in terms of the detailed distribution of facies and distinct definition.

Acknowledgements

We wish to thank Equinor and partners for access to the Volve 3D seismic data used in this exercise. We would also like to thank an unknown reviewer for painstakingly looking through the manuscript and making suggestions. Finally, we thank Gwenola Michaud, who chairs the *First Break* Editorial Board, for making some suggestions that improved the quality of the manuscript.

References

Barnes, A.E. [2007]. Redundant and useless seismic attributes. *Geophysics*, 72, P33-P38.

Bishop, C.M., Svensen, M. and Williams, C.K.I. [1998]. The generative topographic mapping: *Neural Computation*. **10**(1), 215-234.

Chopra, S., Lubo-Robles, D. and Marfurt, K.J. [2018]. Some machine learning applications in seismic interpretation. *AAPG Explorer*, 22-24.

Ha, T., Lubo-Robles, D., Marfurt, K.J. and Wallet, B.C. [2021]. An in-depth analysis of logarithmic data transformation and per class normalization in machine learning: Application to unsupervised

classification of a turbidite system in the Canterbury Basin, New Zealand, and supervised classification of salt in the Eugene Island mini-basin, Gulf of Mexico, manuscript accepted for publication in Interpretation.

Honorio, B.C.Z., Crus Sanchetta, A., Pereira Leite, E. and Campane Vidal A. [2014]. Independent component spectral analysis. *Interpretation*, 2, SA21-SA29.

Hyvärinen, A., and Oja, E. [2000]. Independent Component Analysis: Algorithms and Applications, *Neural Networks,* **13**(4-5), 411-430.

Kohonen, T. [1982]. Self-organized formation of topologically correct feature maps: *Biological Cybernetics*. 43, 59-69.

Kohonen, T. [2001]. Self-organizing Maps: Springer-Verlag.

Lubo-Robles, D. [2018]. Development of independent component analysis for reservoir geomorphology and unsupervised seismic facies classification in the Taranaki Basin, New Zealand: M. Sc. thesis, University of Oklahoma.

Roy, A. [2013]. Latent space classification of seismic facies: PhD Dissertation, The University of Oklahoma.

Roy, A., Romero-Peleaz, A.S., Kwiatkowski, T.J. and Marfurt, K.J. [2014]. Generative topographic mapping for seismic facies estimation of a carbonate wash, Veracruz Basin, southern Mexico, *Interpretation*. 2, SA31-SA47.

Sen, S. and Ganguli, S.S. [2019]. Estimation of pore pressure and fracture gradient in Volve Field, Norwegian North Sea. SPE-194578-MS.