# Mounds labeled data expansion in full-azimuth seismic data via Bayesian generative models

Pedro Silvany* (Petrobras), Ariely Luparelli (Petrobras), Marcos Machado (Petrobras)

This paper was prepared for presentation during the 18th International Congress of the Brazilian Geophysical Society held in Rio de Janeiro, Brazil, 16-19 October 2023.

## Abstract

The static and dynamic data shortage in most subsurface reservoirs makes it extremely difficult to understand the spatial distribution of the different scales of geological heterogeneities (e.g., mounds, natural fractures). The absence of labeled data complicates the application of supervised machine learning algorithms. So, curating labeled training data has become the primary bottleneck in machine learning. Deep generative models have been proposed to synthesize labels at a proper scale in areas with weak supervision sources to accomplish the labeled data scarcity problem. Recently, a joint labeled seismic data expansion generative method has been proposed based on Variational Autoencoders (VAE) and Gaussian Mixture Models (GMM). In this paper, we extend this strategy by using as input multi-channel data (pre-stack full-azimuth seismic data) as considering independent channels by azimuths. Moreover, a Bayesian Gaussian Mixture Model (BGMM) prior conditioned by the variational inference is proposed to fit the deep feature distribution of each label class. The probabilistic Gaussian mixture model is resampled for its class to provide depth features expansion into the decoder and generate expansion-labeled seismic data. This strategy is applied to a Santos basin Pre-salt reservoir to expand labeled mounds facies identified by wells. The approach quickly overcame an important labeled data issue to support seismic characterization, minimizing overfit problems and improving the recognition of mounds' architectural elements in the field.

## Introduction

The static and dynamic data shortage in most subsurface oil and gas reservoirs makes it extremely difficult to understand the spatial distribution of the different scales of geological heterogeneities and, consequently, influences in obtaining realistic flow scenarios. 1D data generally consists of indirect data from electric profiles or direct data acquisition from cores and side samples; both are acquired sparsely (Nelson, 2001). When using only well log data or rock samples, the natural heterogeneous behavior of the geological patterns can be incorrectly represented both by the sampling bias in the well, caused by the well orientation concerning the average direction of the structures, and by the uncertainty in the extrapolation between the wells (Lorenz & Hill, 1994). To fill a 3D reservoir model, we must define field-scale characteristics between wells.

Machine learning algorithms have been widely used in subsurface reservoir modeling to bring quantitative criteria based on seismic data for defining the parameterization and spatial distribution of the main geological features within a production zone (Li et al., 2021; Liu et al., 2023). Nevertheless, the need for labeled data (prior knowledge) complicates implementing supervised strategies. Curating labeled training data has become the primary bottleneck in supervised machine learning due to the scarcity and bias of labeled data from wells. In some specific deep learning tasks (e.g., seismic pattern recognition), it is necessary having thousands or more labeled data (Bach et al., 2017). Moreover, producing labels can be overwhelming due to the specialized domain expertise required (Eadicicco, 2017). To overcome this bottleneck, the adoption of generative models for synthesizing training data from weak supervision datasets has spread to subsurface reservoir modeling.

Deep Neural networks are one of the many methods to obtain a function approximation due to the ability to learn representations. For example, Li et al. (2020) proposed a semi-supervised methodology combining Variational Autoencoders (VAE) with the Gaussian Mixture Model (GMM) to expand post-stack seismic labeled data. The VAE algorithm is an unsupervised generative model that learns efficient data encodings from the seismic data as input (Kingma et al., 2013). Furthermore, unlike an autoencoder (AE) algorithm, the VAE forces the latent variables coded to become normally distributed, making the latent space more continuous and less sparse (Higgins et al., 2021), which brings some benefits. For instance, the variability observed in the interpolated latent space can be used to synthesize new data.

To tackle the over-fitting problem, the VAE training considers the whole seismic data as input. Once VAE is trained, the Encoder architecture encodes each labeled data separately, regarding the wells prior knowledge to obtain the labeled deep feature distribution. Then, the labeled deep feature codes are submitted to a GMM to fit the data distributions and obtain probability density functions for each class. Next, despite the clustering purpose, we expand the labeled depth features by resampling the probability density functions models to obtain pseudo-labeled deep features codes. Finally, the resampled pseudo-labeled codes are submitted to the Decoder architecture to generate new pseudo-labeled samples.

This paper expanded the Li et al., 2020 methodology to overcome a Mound labeled data weak supervision in a pre-salt reservoir from Santo's basin, Southeast Brazilian margin. The methodology considers the possibility of using jointly different types of input data arranged as different channels of the VAE input, such as full-azimuth pre-stack seismic data, as shown by Silvany et al., 2021. The Bayesian VAE-BGMM semi-supervised labeled data expansion method overcomes an important issue for subsurface reservoir characterization due to the unbalanced learning problems (e.g., mounds features, natural fractures zones) and allowing supervised methodologies applications.

## Reservoir Geological Setting

The field is located in the Pre-salt Santos basin, Brazilian Southeast margin. The rocks correspond mainly to the lacustrine carbonates of the Barra Velha Formation, deposited during the Aptian age and, subordinately, to the bioclastic carbonates (coquinas) of the Itapema Formation of the Barremian age (Figure 1). The structural configuration is the dominant factor in controlling the distribution of the facies and the characteristics of the identified reservoirs since it directly influences the dynamics of the circulation of the Aptian/Barremian lake. The reservoir would have remained as a structural high throughout its deposition, conditioned by the activity of normal faults, with a preferential direction N-NW to N-NE resulting from overlapping deformational events related to the predominantly extensional kinematics of the rift tectonics.

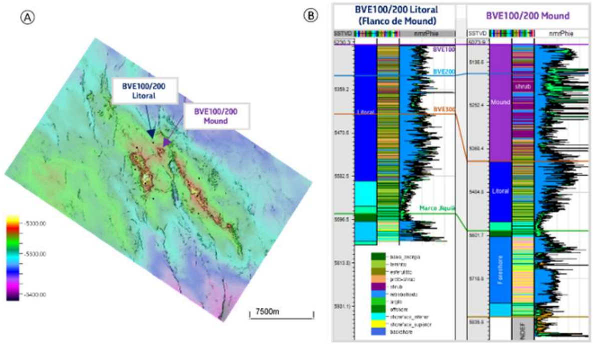


**Figure 1** (a) Structural map of the salt base of the field. Note the occurrence of Mounds features with preferential N-NW to N-NE direction associated with the BVE100 and BVE200 zones. (b) The correlation between wells A and B shows the main depositional facies for the BVE100 and 200 reservoir zones.

The reservoir is marked by carbonate mounds composed of precipitated and/or travertine facies of excellent permo-porous quality, associated with reworked facies, characterized mainly by intraclast grainstones and rudstones (Figure 1). The latter assumes particular importance in the southeastern region of the field due to the presence of an elongated horst in the NW direction, which would have favored sedimentation in a context of high energy. Instead, towards the flank of the mound structures and the main horst, facies developed in an environment of low depositional energy predominate, comprising spherulites and laminites, which may have a high proportion of magnesian clay and organic matter. In addition to depositional aspects, secondary features resulting from dissolution, brecciation, dolomitization, and silicification processes are observed, attributed to hypogenic karstification, with strong control of faults and fractures in the percolation of diagenetic fluids (Figure 1).

## Variational Autoencoders

To apply an auto-codifier network for generative purposes, we must be sure that the latent space is somehow regular (Doersch, 2016). So, to parametrize coding function $f_\varphi$, it will be used an Auto-Encoding Variational Bayes (AEVB) approach, which uses a Stochastic Gradient Variational Bayes (SGVB) estimator to approximate a posterior inference and generate a continuous latent space (Kingma et al., 2013; Doersch, 2016). The process consists of generating a prior distribution $P*\theta(Z)$ and a value Xi from some prior conditional distribution $P*\theta\ (X|Z)$. We assume that the prior and likelihood come from parametric families of distributions $P*\theta(Z)$ and $P*\theta(X|Z)$ and that their PDFs are differentiable almost everywhere (Kingma et al., 2013).
The VAE architecture is composed of two subnets (Figure 2). The encoder will refer to the probabilistic recognition model $f_\varphi(z|x)$ since, given a datapoint $X_i$ it produces a distribution (e.g., a Gaussian) over the possible values of the code $Z_i$ from which the datapoint $X_i$ could have been generated. The encoder subnet is composed of a sequence of convolutional and max pooling layers that are applied over the input as a convolution, which permits identifying patterns in the input image in a way that is invariant with translation. The max pooling layer down samples the input. The encoded distributions are chosen to be normal so that the encoder can be trained to return the mean ($Z_{mean}$) and the covariance ($Z_{dev}$) matrix that describe these Gaussians, which are used to obtain the latent feature $Z = Z_{mean} + Z_{dev}$ for each input $X_i$ (Figure 2).
On the other hand, the decoder will refer to a probabilistic reconstruction model $f_\theta(x|z)$ since, given a code $Z_i$ it produces a distribution over the possible corresponding values of $X_i$, as similar as possible to the original seismic input (Figure 2). The decoder subnet implements an inverted pyramid: it is composed of a sequence of upsampling and convolution layers. Upsampling layer typically doubles the size of the image, assigning to the output pixel the nearest pixel of the input.

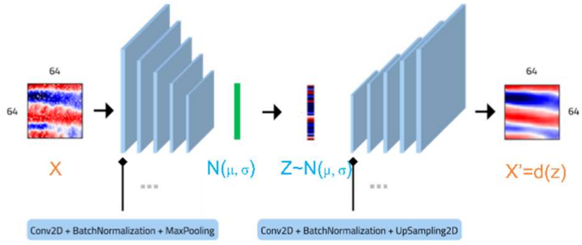

64
64
X
N(μ, σ)
Z~N(μ, σ)
X'=d(z)
64
64
...
...
Conv2D + BatchNormalization + MaxPooling
Conv2D + BatchNormalization + UpSampling2D

**Figure 2** The structure of VAE.

The training of the VAE does not depend on labelled data, it is an unsupervised learning method. The learning is done by minimizing the evidence lower bound (ELBO) objective function (equation 1) (Kingma et al., 2013). The ELBO considers two terms: the first term measured the differences between the posterior and prior gaussian distributions using Kulback-Leibler divergence (KL). This term regularizes the organization of the latent space by making the distributions returned by the encoder $f_\varphi(z|x)$ as close as possible to a standard normal distribution $p_\theta(z)$. The second term measure the reconstruction error for an input $X_i$, where an encoding Z is sampled from $f_\varphi(z|x_i)$, then the probability density of a perfect reconstruction is given by $p_\theta(x_i|z)$.

$\mathcal{L}\ (\theta,\ \phi;\ x) = -DKL(q_\phi(z|x)||p_\theta(z)) + Ez{\sim}q_\phi[\log p_\theta(x|z)]$ (eq. 1)

## Gaussian Mixture Models

Clustering algorithms, as K-means (Lloyd, 1982) or self-organizing map (SOM) (Kohonen, 1990) are based on the notion of distance or dissimilarity. Although, those approach have limitations in estimate the uncertainty measure or probability, which defines the reliability that a data point is associated with a specific cluster. Instead, finite mixture methods as Gaussian Mixture Models (GMM) attempt to do it (McLachlan et al., 2000).

The GMM is a function that is comprised of several Gaussians, each one associated with a cluster. Each Gaussian in the mixture is comprised of the mean (μ) that defines its center, a covariance (Σ) that defines its width and (iii) a mixing probability (π) that defines how big or small the Gaussian function will be (Figure 3). Each Gaussian explains the data points $Z_i$, and the mixing coefficients are themselves probabilities.

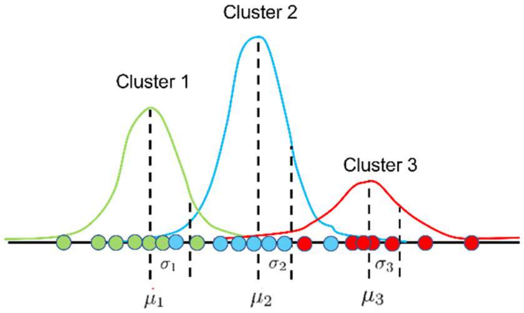


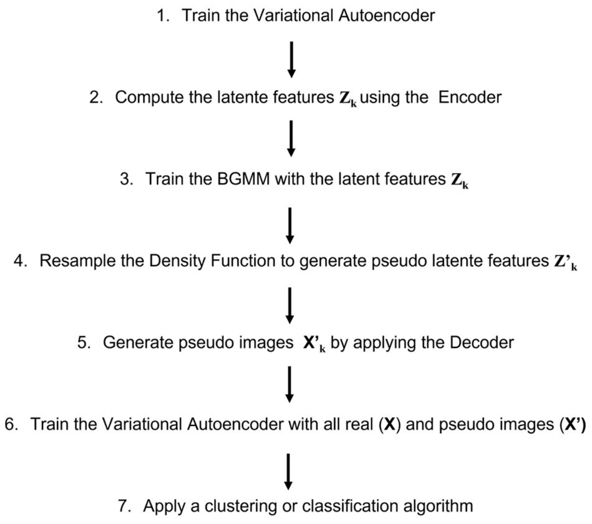
**Figure 3** Image showing three Gaussian functions', hence $K$ = 3.

Additionally, a variational Bayesian approach can be used to fit the GMM, named Bayesian GMM (BGMM). Each point parameter of our model is a variational posterior probability distribution (Roberts et al., 1998). The training to optimize the parameters of those variational posterior distributions to be as close as possible to the true posteriors is done with stochastic variational inference.

## Method

Li et al, 2020, proposed that once trained the VAE in the entire data set $X \subset \Re^{mxN_1xN_2}$, we will transform the labeled data $\mathbf{X_k}$, $X \subset \Re^{mkxN_1xN_2}$with the nonlinear mapping $f_\theta$: $\mathbf{X_k} \rightarrow \mathbf{Z_k}$, where m is the number of samples, k is the index category, $N_1$ is the number of time samples, $N_2$ is the number of traces samples. We use VAE to parametrize the transformation function $f_\theta$. Then, we apply the trained Encoder to make the transformation $f_\theta$: $\mathbf{X_k} \rightarrow \mathbf{Z_k}$ for the input panels considering the prior K labeled data. Observe that the input for the Encoder are the same data used in the VAE training, but now separated by the prior index category. The coded set $\mathbf{Z_k}$ is submitted to the BGMM clustering algorithm to fit the probability density functions for each class $f(z|\mu_{k,}\ \Sigma_k)$. Also, the Encoder outputs $Z_{kmean}$ and $Z_{kdev}$ are used to condition the BGMM prior parameters. After that, we resampled the BGMM probability density function separately by the index category k to obtain the pseudo deep features $\mathbf{Z'_k}$. The final step is input the pseudo deep features $Z'_k$ in the Decoder to generate pseudo labeled samples $X'^k$. Henceforth the VAE training will be done using all real and pseudo images to overcome unbalanced learning problems or allow supervised methodologies applications.

Figure 4 shows the methodology applies in this paper.



1. Train the Variational Autoencoder
2. Compute the latente features $\mathbf{Z_k}$ using the Encoder
3. Train the BGMM with the latent features $\mathbf{Z_k}$
4. Resample the Density Function to generate pseudo latente features $\mathbf{Z'_k}$
5. Generate pseudo images $\mathbf{X'_k}$ by applying the Decoder
6. Train the Variational Autoencoder with all real (**X**) and pseudo images (**X'**)
7. Apply a clustering or classification algorithm

**Figure 4 -** The methodology.

## Pre-salt reservoir case study

The results for the Pre-salt reservoir were obtained with the analysis of full-azimuthal broadband seismic data. The data was obtained by reprocessing using two available seismic data acquisitions. The first is a 3D seismic survey carried out in 2010, when 77 circles with a radius of 6.25 km were acquired, making up an area of 222 km² of azimuthal coverage greater than 180°. In addition, a second 3D streamer acquisition was included, heading N122°, the most recent one in the area. The objective of the reprocessing was to improve the imaging of the area of interest, with high structural and stratigraphic complexity, in addition to providing anisotropy information and the possibility of exploring full-azimuthal attributes for the characterization of the reservoir.

Pre-stack seismic gathers and CWT Voices attributes (Silvany et al., 2021) referring to four azimuth sectors (N0o, N45o, N90o, and N135o) were collected to extract time-offset and time-frequency panels as a multi-channel input for training the VAE network. The BVE110 and BVE200 intervals were chosen. The BVE110 zone has an average thickness of 95 meters, with a maximum of 292 meters in the carbonate mounds. The BVE200 zone has an average thickness of 110 meters, with a maximum of 310 meters in the carbonate mounds area. The time-offset and time-frequency panels both have 25 traces. The feature space is generated as a vector space with a dimension of 32 components, equivalent to a dimensionality reduction of approximately 95%. Figure 5 shows examples of the input data, the coded latent features, and the network reconstructions by azimuth.

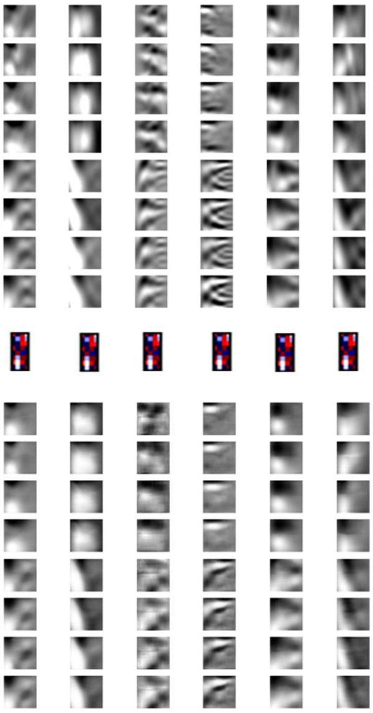


**Figure 5** (a) Six examples of input time panels extracted from CDP's and Voices collections for azimuths N0o, N45o, N90o, and N135o, respectively; (b) Six examples of encoding generated by VAE; (c) Six examples of reconstructed outputs images.

Figure 6a shows the mounds prior seismic labeled data occurrence for the BVE reservoir zones in red, regarding the well's cores and logs interpretation, and cautions rifts carbonate depositional conceptual model extrapolation (Figures 1). However, those prior areas constitute less than 2 percent of the whole seismic input data and, given the small data quantities, constitute an unbalanced problem for mound deep-learning architecture learning and prediction far from the wells. Therefore, we introduce the VAE-BGMM methodology to add new pseudo azimuthal pre-stack images as input for the VAE training. Figure 7a and 7b shows the pseudo-reconstructed images **X'** and pseudo deep codes **Z'**, respectively.

First, the whole pre-stack and voice azimuth data were introduced to the VAE-BGMM architecture learning. Figure 6b shows the seismic facies map with six clusters obtained by applying the methodology only considering the real data. Then, once VAE was trained, new pseudo azimuthal pre-stack mounds images were generated by decoding the pseudo-latent features **Z'ₖ** (Figure 7). Hence, the pseudo-new images were encoded with the original seismic data resulting in the seismic facies map with six cluster shown in Figure 6c. Notice that, despite the similarities in both maps (Figure 6b and 6c), the mounds occurrence zones (purple color) in Figure 6c are more restricted regarding the mounds conceptual model for the area and have sharp geometries. Nevertheless, both maps constitute different scenarios that shall contribute for the depositional facies model understand.

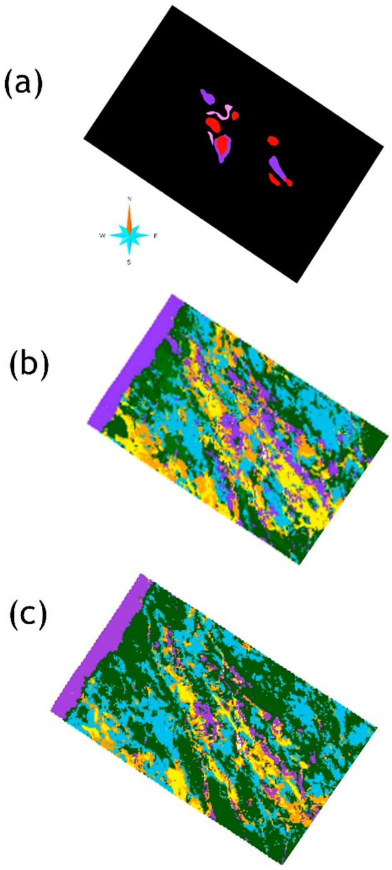


**Figure 6** (a) Mounds prior seismic labeled data occurrence for the BVE reservoir zones; (b) Map of six seismic facies corresponding to the BVE100 zone by applying the VAE in real input data; (c) Map of six seismic facies corresponding to the BVE100 zone by applying the VAE considering pseudo mounds expanded images.

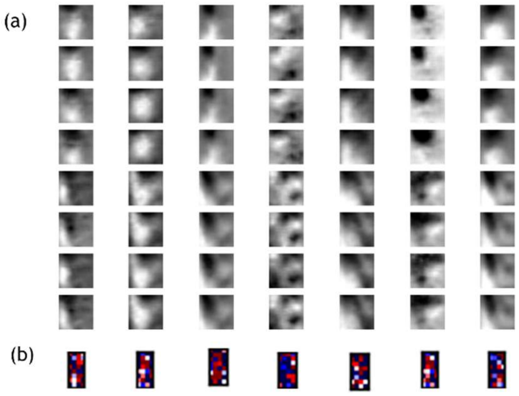


**Figure 7** (a) Mounds pseudo reconstructed images **X'**; (b) Mounds pseudo deep codes **Z'**.

The seismic facies map obtained with the VAE-BGMM expanded methodology showed good correspondence with the information from the impedance cube P (Figure 8), which, in turn, had the curves of the ten wells drilled in the field as input.

One possible interpretation for the seismic facies map is that the shades of purple are associated with regions of better permo-porosity, attributed to mound features or reworked facies. These deposits are associated with (1) Mounds deposited in high energy environment with chemical precipitation and growth of the sedimentary section, with a predominance of shrubs/stromatolites and reworked facies of moderate to high energy (intraclasts grainstone/rudstone); (2) Shallow Plains associated with structural highs with higher proportions of reworked facies of moderate to high energy (intraclasts grainstones/rudstones) and development of shrub fascicular crusts or shrubs/stromatolites (ETR), and (3) Coast subenvironment with a predominance of facies with good permo-porous characteristics, shrubs/stromatolites and reworked (intraclasts grainstones/packstones), interspersed with spherulitic levels (Figure 8). Furthermore, the orange and yellow colors were related to low energy environments, where spherulites and laminites would predominate, exhibiting less permo-porous quality; these attributed to the (4) Sublittoral. Finally, the shades of green evidence a degradation of the reservoir, with a predominance of laminated facies of low energy, constituting the non-reservoir rocks of the (5) Protected Environment or of (6) Deep Lake predominantly.

Based on the seismic facies map, interpretation of the pattern of reflectors, and the extrapolation from the wells, the reservoir geological model was subdivided in those detailed depositional subenvironment (Figure 8e).

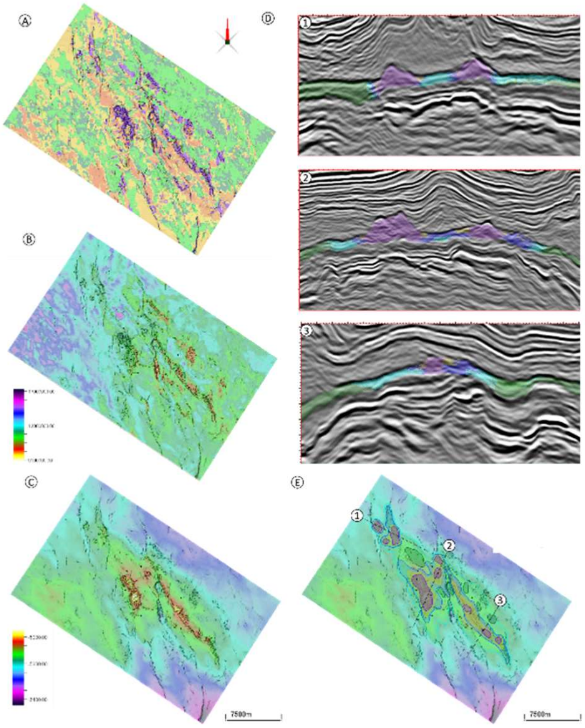


**Figure 8** (a) Map of six seismic facies corresponding to the BVE100 zone by applying the VAE considering pseudo mounds expanded images; (b) Mean acoustic impedance map for the BVE100 range; (c) Structural topo map of the reservoir; (d) Amplitude volume sections with subenvironments defined for be; (e) Structural topo map of the reservoir, with the polygons of subenvironments mapped to this stratigraphic interval. Dashed lines correspond to the location of the sections shown in d. All maps are superimposed on the similarity attribute.

## Conclusions

The VAE-BGMM methodology allowed us to generate new pseudo-labeled data by tackling the problem of latent space irregularity instead of a single point, ensuring a better organization of the latent space. Furthermore, the strategy overcomes unbalanced learning problems and allows apply supervised methodologies in weak supervised recognition areas. Comparing the seismic facies results with the information obtained with inversion attributes allowed greater detail, reliability, and robustness in the geophysical mapping. In addition, frequency data obtained by CWT Voices improved the recognition of small-scale structural and stratigraphic heterogeneities. Both results must constitute different scenarios that contribute to understanding the depositional facies model.

## Acknowledgments

We thank Petrobras for permission to publish this work.

### References

Bach, S. H., He, B., Ratner, A., & Ré, C. (2017, July). Learning the structure of generative models without labeled data. In *International Conference on Machine Learning* (pp. 273-282). PMLR.

Doersch, C. (2016). Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*.

Eadicicco, L. Baidu's Andrew Ng on the future of artificial intelligence, 2017. *Time [Online]*.

Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., & Botvinick, M. (2021). Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications*, *12*(1), 6456. DOI: https://doi.org/10.1038/s41467-021-26751-5.

Kohonen, T., 2001, Self-organizing maps: Springer.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
Kohonen, T., 2001, Self-organizing maps: Springer. https://doi.org/10.48550/arXiv.1312.6114.

Li, K., Chen, S., & Hu, G. (2020). Seismic labeled data expansion using variational autoencoders. *Artificial Intelligence in Geosciences*, *1*, 24-30. https://doi.org/ 10.1016/ j.aiig.2020.12.002.

Lloyd, Stuart P. "Least squares quantization in PCM." Information Theory, IEEE Transactions on 28.2 (1982). **DOI:** 10.1109/TIT.1982.1056489.
Li, Y., Wang, C., Tian, Y., & Wang, S. (2021). Parameter-shared variational auto-encoding adversarial network for desert seismic data denoising in Northwest China. *Journal of Applied Geophysics*, *193*, 104428. https://doi.org/10.1016 /j. jap pgeo.2021.104428
Liu, Y., and I. Tsvankin, 2021, Methodology of time-lapse elastic fullwaveform inversion for VTI media: Journal of Seismic Exploration, 30, 257–270.
Liu, Y., Feng, S., Tsvankin, I., Alumbaugh, D., & Lin, Y. (2023). Joint physics-based and data-driven time-lapse seismic inversion: Mitigating data scarcity. *Geophysics*, *88*(1), K1-K12. https://doi.org/10.1190/geo2022-0050.1

Lioyd, Stuart P. "Least squares quantization in PCM." Information Theory, IEEE Transactions on 28.2 (1982).

Lorenz, J. C., & Hill, R. E. (1994). Subsurface fracture spacing: Comparison of inferences from slant/horizontal and vertical cores. *SPE Formation Evaluation*, *9*(01), 66-72. https://doi.org/10.2118/21877-PA .

McLachlan, G., & Peel, D. (2000). Finite mixture models: Wiley Interscience.

Nelson, R. (2001). *Geologic analysis of naturally fractured reservoirs*. Elsevier.

Roberts, S. J., Husmeier, D., Rezek, I., & Penny, W. (1998). Bayesian approaches to Gaussian mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1133-1142. **DOI:** 10.1109/34.730550.

Silvany, P. H., Machado, M., Matos, M., & Paes, M. (2021, September). Joint multiazimuthal prestack and time-frequency attributes seismic facies prediction via deep learning: An application to a Brazilian presalt reservoir. In *First International Meeting for Applied Geoscience & Energy* (pp. 1480-1484). Society of Exploration Geophysicists. https://doi.org/10.1190/segam2021-3594815.1.

Viroli, C., & McLachlan, G. J. (2019). Deep Gaussian mixture models. *Statistics and Computing*, *29*, 43-51. DIO: https://doi.org/10.1007/s11222-017-9793-z.