Different training sample selection strategies in unsupervised seismic facies analysis
Tao Zhao* and Kurt J. Marfurt, University of Oklahoma

SUMMARY

Pattern recognition based multiattribute seismic facies analysis enables seismic interpreters to effectively extract and analyze information buried in several seismic attributes. However, most pattern recognition methods rely heavily on training data, which means the algorithms detect features that best represent the training data. In this study, using self-organizing map as an example of pattern recognition techniques, we discuss the influence (and sometimes, bias) associated with different training data selection strategies. We further demonstrate that using the same attributes, different training samples may lead to different interpretation of seismic facies.

INTRODUCTION

Recent development in seismic facies analysis techniques permits interpreters to simultaneously analyze multiple seismic attributes, well logs, as well as data from other disciplines. One type of the popular seismic facies analysis techniques is unsupervised learning, which uses clustering or projection to automatically discover natural seismic facies residing in several seismic attributes. Commonly, interpreters take data from a window of interest, build a model of clusters using all or part of the data within the analysis window (these are the training data), then apply the model, assigning clusters to every data sample within the window. Using such techniques requires special care when selecting training data, because such training data determine the unsupervised learning model, and the model determines the seismic facies.

Different training data lead to different seismic facies maps. Often time, interpreters spend much of their effort on selecting the suite of attributes to use as well as testing parameters of the learning algorithm, but care less about where the training samples come from the \((x, y, z)\) space spatially. Commonly, interpreters use all data samples either volumetrically (Roden et al., 2015), or along a horizon (Zeng, 2004). Because of the size of modern 3D seismic data, an alternative and similar approach is to uniformly sample the input data. Roy et al. (2014) use uniform subsampling to reduce the computation cost of the generative topographic mapping (GTM) algorithm to a manageable size when characterizing a carbonate wash reservoir in Mexico. When interpreters are interested in features from a particular region and want to explore the presence of such features in a broader area, picking training data from that particular area will help to capture the features of interest. In Roy et al. (2014), the authors also use training data extracted around wellbores to build a GTM model and map the extension of facies found at well locations. However, until today, there are very few, if any published studies discussing the bias towards more dominant features if the feature of interest composes a much smaller population in the input data. Examples include bright/dim spots, mid-channel bars, and gas-hydrates. Most unsupervised seismic facies analysis algorithms look for principal trends but omit small populations (e.g. principal component analysis based methods). We show an example of how a small population collapses into the main population in Figure 1. Such phenomenon encourages us to explore different training sample selection strategies if such small population features are the exploration target.

Figure 1. In the case of two attributes, a feature of small population (green dots) collapses into the main population when projecting onto the first eigenvector.

Self-organizing map (SOM) is one of the most popular unsupervised learning algorithms that has been successfully applied to seismic interpretation problems. The very first applications of SOM on seismic data include Strecker and Uden (2002), in which the authors use multiattribute SOM volumetrically, and Coleou et al. (2003) use both seismic amplitudes (waveform classification) and seismic attributes as inputs for SOM. Zhao et al. (2016) and Zhao et al. (2017) introduce distance preservation during projection and stratigraphic constraint to improve the SOM performance on seismic attribute data.

In this study, we use SOM as an implementation of unsupervised learning method, and experiment with four
different training sample selection strategies, namely uniform sampling, user-guided sampling, weighted sampling, and reflector density based sampling. We will briefly introduce the frameworks of these four different strategies in the next section, and will then compare the facies map generated from training sets selected using the four approaches.

**TRAINING SAMPLE SELECTION STRATEGIES**

**Uniform sampling**
Uniform sampling is the most straightforward training sample selection approach, and one extreme case uses all data samples to train the model. Typically, we select one sample for every $X$ line/cdp and $Y$ time/depth samples. A schematic plot for uniform sampling is shown in Figure 2.

![Figure 2](image.png)

*Figure 2.* The schematic plot for uniform sampling. Red dots are sampled training data.

**User-guided sampling**
We loosely define user-guided sampling as extracting training data from a user-defined region about a feature of interest. Common practices that use a time window or window confined by two horizons also fall into this category; yet in this study, we focus the user-guided sampling on arbitrary shaped regions with an objective to recover the details in a turbidite channel complex. A schematic plot for user-guided sampling is shown in Figure 3, with a field data example in Figure 4.

![Figure 3](image.png)

*Figure 3.* The schematic plot for user-guided sampling. Two polygons are regions to sample training data.

![Figure 4](image.png)

*Figure 4.* A field data example for user-guided sampling. Magenta points are training samples selected within a channel complex.

**Weighted sampling**
In case that the interpreter’s purpose is to recover details in a small region while still producing the main facies in the analysis window, he/she may want to draw training samples uniformly from the whole analysis window and combine with samples from the small region of interest. The interpreter may further want the model to favor samples associated with the small region of interest. By assigning different weights to samples within or outside the region of interest, weighted sampling is a means to emphasize the subtle facies of small population. A schematic plot for weighted sampling is shown in Figure 5.

![Figure 5](image.png)

*Figure 5.* The schematic plot for uniform sampling. Red dots are sampled training data with lower weight. Samples within the two polygons are of higher weight.

**Reflector density based sampling**
We assume that the degree of geology change is proportional to the reflector density. To reduce the number of redundant samples that represent gently changing regions and increase samples representing rapidly changing regions, we propose to sample more in regions of denser reflectors. We use a 3D Sobel filter to highlight the edges between peak and trough reflectors, then use a mean filter to generate a pseudo reflector density map, and finally use a weighted random selection to extract training samples based on the reflector density map. Where the reflector density is higher, it has a higher probability to be selected, and vice versa. Figure 6 summarizes the workflow.

**APPLICATION ON A FIELD EXAMPLE**
We use the four different training sample selection schemes on a data set from Canterbury Basin, offshore New Zealand. The objective is to characterize a turbidite channel system, and we expect to observe various degree of details from different training sets. Sutherland and Browne (2003) provide details on the depositional history and hydrocarbon potentials. A more detailed description regarding the turbidite channel system is provided in Zhao et al. (2016), and in this study, we follow their workflow and focus on the influence of different training sets. The input attributes for SOM algorithm are peak spectral frequency and magnitude, coherent energy, and GLCM homogeneity. The attributes are sensitive to channel thickness, lithology, and subtle geomorphology changes. The number of training samples used in each test is listed in Table 1.
Facies analysis training sample selection

User-guided sampling result is shown in Figure 8. We pick training samples specifically around the two channels marked by the orange and red arrows, and by doing so, we are able to identify the subtle changes marked in Figure 11 that are previously not identified in Figure 7. On the other hand, because the training samples are selected around the two channels, the colors at regions outside this relatively small area do not necessary represent the real facies. Instead, they only represent the similarity between a facies and facies within the dual-channel region.

Figure 9 gives the result from weighted sampling, in which we take samples from uniform sampling and user-guided sampling, and assign weight 1 for uniformly sampled points and 2 for user-picked picked points. By combining the two sampling schemes, we are able to observe some of the subtle changes within the dual-channel system (e.g. the edge between two channel stories marked by the blue arrow). At the same time we still preserve the meaning of the facies outside the dual-channel region, because the majority of training samples are extracted outside the dual-channel region.

Finally, Figure 10 gives the result from reflector density based sampling. Being our first attempt, the result from reflector density based sampling provides similar information as the weighted sampling. However, it looks noisy and contains orange colored patches that are not geologically plausible. We attribute this less than optimal performance to the fact that the reflector density calculation is overly sensitive to high amplitudes. All high amplitude regions are characterized by the orange color that biases the whole facies map.

CONCLUSION

In this study, we explore the influence of different training sampling selection schemes. By comparing the SOM facies maps from four different training sets, we observe that uniform sampling is the best for discovering the main facies, and user-guided sampling preserves the details within a small region. Weighted sampling provides an in-between solution, for which the ratio and weight between uniformly sampled points and user-picked points need to be carefully defined. Reflector density based sampling is promising but needs to be independent from reflector amplitudes.

ACKNOWLEDGEMENT

We thank New Zealand Petroleum and Minerals for providing the seismic data to public. Financial support for this effort is provided by the industry sponsors of the Attribute-Assisted Seismic Processing and Interpretation (AASPI) consortium at the University of Oklahoma. We thank colleagues for their valuable input and suggestions. The horizon slice visualization is performed using licenses to Petrel, provided to the University of Oklahoma for research and education courtesy of Schlumberger.
Facies analysis training sample selection

Figure 7. SOM facies from uniform sampling along Horizon A. Inserts aa', bb', and cc' are given in Figure 11.

Figure 8. SOM facies from user-guided sampling along Horizon A. Inserts aa', bb', and cc' are given in Figure 11.

Figure 9. SOM facies from weighted sampling along Horizon A. Inserts aa', bb', and cc' are given in Figure 11.

Figure 10. SOM facies from reflector density based sampling along Horizon A. Inserts aa', bb', and cc' are given in Figure 11.

Figure 11. Seismic amplitude along vertical inserts aa', bb', and cc'. The red curve denotes Horizon A. Yellow arrow marks a channel boundary, blue arrow an edge between two channel stories, and white arrows changes in reflection characteristic.

---

© 2017 SEG
SEG International Exposition and 87th Annual Meeting
REFERENCES


