Papers with Code: Unsupervised Seismic Waveform Classification

Author: Roderick Perez Altamar

ABSTRACT

Unsupervised waveform classification is a proven technology for objective seismic facies analysis, yet its application often remains confined to proprietary commercial software. This tutorial bridges that gap by presenting a complete, hands-on workflow that empowers geoscientists to implement this technique using open-source Python tools. The methodology is built around the K-Means clustering algorithm and is demonstrated through a progressive series of examples: a foundational 2D synthetic model, a more complex 2.5D meandering channel system, and a final application to a real-world 3D seismic volume.

Readers will learn the essential, practical steps to independently execute the entire analysis, from extracting waveform data around a horizon to determining the optimal number of facies with the Elbow Method and visualizing the final, interpretable maps. The key takeaway is the ability to move beyond being a user of black-box software to becoming a creator of customized analytical workflows. This guide provides the practical knowledge needed to apply a powerful, data-driven interpretation technique, enabling greater flexibility, understanding, and innovation without the need for a commercial license.

INTRODUCTION

Seismic exploration ultimately seeks to construct a geologically consistent reservoir model that captures both the spatial distribution of petrophysical properties and the heterogeneities that govern fluid flow. Because the recorded seismic signal arises from reflectivity contrasts between subsurface layers, amplitude variations within a seismic trace commonly reflect changes in depositional environment, lithologic composition, or fluid saturation. When those variations are examined collectively, the resulting seismic character can be interpreted in terms of seismic facies, providing quantitative insight into stratigraphic architecture and sedimentological processes.

A seismic wavelet—the compact pulse emitted or encoded by the source and recorded by the receivers—forms the fundamental building block of every seismic-processing workflow (Enders and Treitel, 2008). Mathematically, a wavelet is a finite-energy signal whose amplitude is largely confined to a limited interval on the time axis (Robinson, 1962; 1964a; 1964b). Because different lithologies and fluid contents modify the propagating wavelet in characteristic ways, its shape has long been exploited to build seismic facies and stratigraphic models (Xu and Haq, 2022) and to pinpoint reservoirs in structurally complex settings (Pico et al., 2019). Today, machine-learning algorithms automate the extraction, clustering, and classification of these wavelets, allowing interpreters to map subtle spatial variations in waveform expression and deepen their understanding of reservoir heterogeneity. Seismic waveforms—or discrete segments of traces—are therefore classified to highlight facies changes within a user-defined interval, whether that

interval spans two distinct geologic formations or simply a constant time or depth window of interest.

Within this interpretive framework, seismic waveform classification stands out as one of the most powerful unsupervised pattern-recognition techniques currently available. By grouping traces with similar shapes and spectral attributes, the method produces purely data-driven facies classes that require no a priori geologic constraint—only the definition of the analysis interval (Andersen and Boyd, 2004). In this tutorial, that capability is demonstrated through a three-stage workflow. First, a synthetic two-dimensional seismic section is generated and convolved with a zero-phase Ricker wavelet to illustrate the foundational concepts. Second, the exercise is extended to a quasi-three-dimensional synthetic volume that emulates a meandering-channel system; random noise is superimposed on the data to mimic the acquisition footprint encountered in field surveys. Finally, a real three-dimensional seismic cube is loaded, a target horizon is picked, and the corresponding wavelets are extracted for classification.

The unsupervised classification itself is performed with the K-means algorithm (MacQueen, 1967). To determine the optimum number of clusters, the elbow criterion is applied to the within-cluster sum-of-squares metric. The resulting class probabilities—or, equivalently, similarity volumes—are rendered as color variations along the reflector of interest, yielding a seismic-waveform distribution map that can be related directly to depositional facies (Priezzhev and Manral, 2012). Because every seismic trace is classified solely on the basis of its intrinsic waveform, the interpreter gains an unbiased view of how seismic character evolves across the survey area and, by implication, how geologic facies vary laterally and vertically.

All code, figures, and intermediate outputs referenced in this article are provided in three self-contained Jupyter Notebooks (Python) hosted at https://github.com/roderickperez/seismicWaveformClassification.git. Readers are encouraged to rerun the notebooks, adjust parameters such as wavelet frequency, noise level, or K-means initialization, and observe firsthand how each modification influences both the synthetic data and the final classification results.

METHODOLOGY

The tutorial unfolds in three progressively richer steps. It begins with a hand-crafted two-dimensional (2D) synthetic section that isolates every processing element, extends those principles into a pseudo-2.5D volume that mimics a migrating meandering channel, and culminates by applying the identical workflow to an open-source 3D seismic dataset (the F3 cube) in which wavelets are extracted along an interpreted horizon. At each stage, the same unsupervised K-means engine is deployed and the elbow criterion is used to confirm the optimum number of waveform classes before any geological interpretation is attempted.

2D Synthetic Example: The Fundamentals

To grasp the core concepts of waveform classification, we begin with a simple, controlled 2D synthetic model. This foundational step involves creating a simplified representation of the Earth's subsurface and simulating the seismic response that would be recorded from it.

The first task is to define a geological model using reflection coefficients (RC). These coefficients represent the acoustic impedance contrasts at the boundaries between different rock layers, governing the strength and polarity of reflected seismic energy. For our model, we construct a 2D grid where a single layer contains a series of laterally changing reflection coefficients, simulating a geological feature of interest, while the surrounding layers are acoustically transparent (Figure 1a). A positive RC value (white) signifies an increase in acoustic impedance (e.g., shale to limestone), while a negative value (black) indicates a decrease.

Next, we must define the source signal, or seismic wavelet, that is theoretically sent into the ground. We use a standard, zero-phase 50 Hz Ricker wavelet, whose characteristic shape is defined by its dominant frequency. The Ricker wavelet is mathematically described as:

$$w(t) = (1 - 2\pi^2 f^2 t^2) e^{-\pi^2 f^2 t^2}$$

where $w(t)$ is the amplitude at time t and f is the dominant frequency (Enders and Treitel, 2008). The shape of this wavelet, with its central peak and side troughs (Figure 1b), is critical as it dictates how geological boundaries will be resolved in the final seismic image. With the geological model and wavelet defined, we generate the synthetic seismic section through a mathematical operation known as convolution. Conceptually, convolution 'stamps' the Ricker wavelet's signature onto each reflection coefficient in our model. This process simulates what an array of geophones would record in a seismic survey, producing a collection of seismic traces that form a 2D seismic section (Figure 1c). In this final image, we can clearly see the 'wiggles' that represent the reflections from our defined geological layer.

Having generated our synthetic data, we can now apply an unsupervised machine learning algorithm to classify the seismic traces. We use K-Means clustering, an algorithm that automatically groups data points—in our case, entire seismic traces—based on their similarity. The goal is to have the algorithm identify and group traces that have a similar waveform shape without any prior geological input.

A crucial parameter for the K-Means algorithm is the number of clusters, or 'facies,' we want to identify. To make an informed decision, we employ the Elbow Method. This technique involves running the K-Means algorithm multiple times with an increasing number of clusters (e.g., from 2 to 10) and plotting a metric known as the Sum of Squared Errors (SSE) for each run. The resulting plot typically forms a curve resembling an arm. The 'elbow' of this curve—the point where the rate of SSE decrease slows significantly—suggests the optimal number of clusters. For our dataset, the elbow clearly appears at k=5, validating this as a suitable choice (Figure 2a).

With the optimal number of clusters determined, we run the K-Means algorithm, which assigns each of the seismic traces to one of the five clusters. To visualize the results, we can first examine

the average, or representative, waveform for each cluster (Figure 2b). These distinct waveform shapes are geologically meaningful; they represent the different seismic facies identified by the algorithm. For instance, some waveforms might represent the seismic response at the core of the geological feature, while others could correspond to its edges or the background geology.

Finally, we can create a seismic facies map by color-coding each trace in the seismic section according to its assigned cluster (Figure 2c). This provides a clear spatial view of the distribution of the different seismic responses. Overlaying this classification map on the original seismic data (Figure 2d) creates a powerful composite image that directly links the data-driven facies to the underlying seismic amplitudes, confirming that the algorithm has successfully delineated areas of distinct seismic character.

2.5D Meandering-Channel Volume

Building on the 2D fundamentals, we now move to a more geologically realistic scenario: a 2.5D model of a meandering river channel. This involves creating a 3D geological volume slice by slice, generating a corresponding 3D seismic cube, and adding random noise to better simulate real-world data acquisition conditions.

The geological model is constructed programmatically, defining parameters such as channel thickness, sinuosity, and lateral position to generate a meandering channel form within a 3D grid of reflection coefficients. To make the classification task more challenging and realistic, we introduce a small amount of random noise to the clean synthetic seismic volume. This step is critical, as real seismic data is always affected by noise from various sources.

Visualizing this 3D data volume is key. We can inspect it through vertical cross-sections (inlines) or horizontal time-slices. Figure 3a shows a horizontal time-slice from the middle of the seismic volume, where the meandering channel is faintly visible amidst the noise. The task for our machine learning algorithm is to enhance the visibility of this feature.

The classification workflow remains similar to the 2D case, but now we apply it to a much larger dataset. The 3D seismic cube is 'flattened' into a 2D table, where each row represents a single seismic trace. The K-Means algorithm is then applied to this table. Again, the Elbow Method is used to determine the optimal number of clusters, which for this dataset is found to be four (Figure 4a).

The results of this classification are compelling. The algorithm successfully groups the noisy traces into four distinct clusters, each with a unique average waveform (Figure 4b). When we reshape the cluster assignments back into a map view, the result is a clear and well-defined image of the meandering channel system (Figure 3b). The algorithm has effectively "seen through" the noise to identify the underlying geological pattern, a task that would be challenging to do manually from the noisy seismic data alone.

To further analyze the result, we can view a multi-panel display for a specific inline (Figure 4c, 4d, 4e). This view shows the original seismic section, the corresponding color-coded cluster map for that line, and a direct overlay of the two. This visualization provides an intuitive link between the raw seismic wiggles and the geological facies they represent, demonstrating how the classification has successfully segregated the channel-related responses from the background seismic character.

3D Field Data: A Real-World Case Study

The true test of any geophysical technique lies in its application to real-world data. In this final section, we apply the unsupervised classification workflow to a 3D seismic survey to identify and map seismic facies around a picked geological horizon.

The first practical step involves loading the 3D seismic data from its industry-standard SEGY file format and the associated horizon interpretation from a CSV file. Specialized Python libraries are used to handle the complexities of SEGY headers, ensuring that inline, crossline, and real-world coordinate information are correctly read and aligned. Figure 5 shows a representative inline from the 3D seismic cube, with the picked horizon highlighted in red. This horizon serves as our guide for the analysis.

Instead of classifying the entire seismic trace, which can be computationally intensive and geologically less specific, we focus our analysis on a small window of seismic data centered on the picked horizon. For this case study, we extract a snippet of the waveform from five samples above to five samples below the horizon for every trace in the survey. These short waveform snippets, which capture the seismic character at and around the target geological interface, become the input for our clustering algorithm.

As before, the Elbow Method is employed to guide our choice for the number of clusters. The analysis suggests that eight distinct seismic facies (k=8) provide a meaningful classification of the waveform variability within the dataset (Figure 6a). Following the K-Means clustering, we compute the average waveform for each of the eight clusters (Figure 6b). This is a critical QC and interpretation step. By examining these representative wavelets, a geoscientist can begin to assign geological meaning to each facies—for example, one cluster might represent a high-amplitude peak characteristic of a hard rock interface, while another might show a low-amplitude, complex response indicative of a transitional boundary.

The final outputs provide a powerful new perspective on the subsurface. We generate a comprehensive seismic facies map (Figure 7a) that displays the spatial distribution of the eight clusters across the entire survey area. This map immediately reveals clear geological patterns, such as channels or fan systems, that were not obvious on a simple amplitude map. The colors delineate distinct regions of seismic character, allowing the interpreter to map geological elements with confidence.

To complete the workflow, we tie this map-view interpretation back to the sectional data. Figure 7b displays the same seismic inline shown earlier (iline 445), but now it is overlaid with colored bars representing the classification results within the analysis window. This provides a direct and intuitive visual link between the raw seismic data, the picked horizon, and the data-driven facies classification, confirming the validity of the interpretation and providing a rich, integrated view of the subsurface geology.

SUMMARY

This tutorial demonstrates a complete and practical workflow for unsupervised seismic waveform classification, progressing systematically from simple synthetic models to a complex, real-world 3D field dataset. The initial 2D and 2.5D synthetic examples served as a validation stage, confirming that K-means clustering can effectively partition seismic traces based on waveform character and identify geological features of increasing complexity, from a simple planar reflector to a noisy, meandering channel system. These controlled tests established the core mechanics of the workflow: forward modeling by convolution, restructuring of the data so that each trace becomes a feature vector, and objective selection of the number of clusters using the elbow criterion.

The true power and applicability of the workflow were realized in the application to the 3D dataset. By integrating a picked geological horizon, the analysis was confined to a stratigraphic interval of interest. Thousands of localized waveform snippets were extracted and clustered, yielding eight distinct seismic facies. The final interpretive products—a detailed wavelet-facies map and representative average waveforms—translate the numerical outcome of the algorithm into tangible, geologically meaningful insight. Because the procedure relies solely on the recorded seismic response, the classifications are entirely data driven; geological meaning must still be supplied by the interpreter, who can relate cluster boundaries to depositional architecture, fluid content, or structural overprint.

Although K-means offers a fast and intuitive entry point, alternative unsupervised techniques—hierarchical clustering, self-organizing maps, or Gaussian-mixture models—can reveal nested or non-spherical cluster structure. Likewise, tying the horizon-centered wavelets back to well control would enable a supervised extension in which known lithologies guide the classifier and sharpen facies predictions along sparsely drilled intervals. Experimenting with different wavelet shapes or frequencies can illuminate the sensitivity of the results to source bandwidth and phase assumptions. Finally, following the recommendation of Andersen and Boyd (2004), blending waveform clusters with complementary seismic attributes and reducing redundancy through principal-component analysis can be particularly powerful when well data are limited, yielding a multi-attribute view of subsurface variability that honors both physics and statistics.

CONCLUSIONS

Unsupervised seismic waveform classification using algorithms like K-Means provides a robust, efficient, and objective method for seismic facies analysis. By progressing from a simple 2D model to a complex, real-world 3D dataset, we have demonstrated how this technique can distill vast amounts of seismic data into interpretable geological maps. This data-driven approach empowers geoscientists to accelerate their interpretation workflows, uncover subtle geological features that might otherwise be missed, and build more confident and detailed models of the subsurface. As computational tools become increasingly accessible, integrating machine learning into the standard geophysical toolkit is no longer a future ambition but a present-day reality that enhances our ability to understand the Earth.

REFERENCES

Andersen, E. and Boyd J., [2004]. Seismic waveform classification: techniques and benefits. CSEG Recorder, 29(3).

Enders, R. and Treitel S. [2008]. *Digital Imaging and Deconvolution: The ABCs of Seismic Exploration and Processing. SEG, Tulsa, 424.*

MacQueen, J. B. [1967]. Some Methods for classification and analysis of multivariate observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 281–297.

Pico, A., Taqi, F., Marzouqpha, A.R.S.R., AlDoub, A.S., Ahmad, A., Al-Dohaiem, K., Tyagi, A., Pabitra, S., Kharghoria, A., Al-Rabah, A. and Zhang, I., [2019]. Characterizing stratigraphic traps using waveform classification of seismic facies: A case study of Shallow Reservoir, Kuwait. In SEG International Exposition and Annual Meeting (p. D033S061R001). SEG.

Priezzhev, I. and Manral, S. [2012]. 3D Seismic waveform classification. In Istanbul 2012-International Geophysical Conference and Oil & Gas Exhibition. Society of Exploration Geophysicists and The Chamber of Geophysical Engineers of Turkey. 1-4.

Robinson, E. A., 1962, Random wavelets and cybernetic systems: Charles Griffin and Co.

Robinson, E. A., 1964a, Wavelet composition of time series, in H. O. A. Wold, ed., Econometric model building, essays on the causal chain approach: North Holland Publishing Co., 37–106.

Robinson, E. A., 1964b, Recursive decomposition of time series, in H. O. A. Wold, ed., Econometric model building, essays on the causal chain approach: North Holland Publishing Co., 111–168.

Xu, G. and Haq, B.U. [2022]. Seismic facies analysis: Past, present and future. Earth-Science Reviews, 224, p.103876.

TABLE OF FIGURES

colored bars overlaid on the seismic data around the horizon. This directly links the map view facies to the raw seismic response.