**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 initial phase of this study focused on validating the core unsupervised classification workflow using a simple, noise-free 2D synthetic model. This approach allowed for the isolation and examination of the fundamental mechanics of the process. The geological substrate was first defined as a two-dimensional digital grid representing a vertical cross-section of the earth. A distinct geological boundary was simulated within this model by defining a horizontal layer with a specific pattern of reflection coefficients (RCs). These coefficients, which ranged symmetrically from -0.9 to 1.0, were designed to model a significant and varied impedance contrast.

To simulate a seismic survey, a standard 50 Hz Ricker wavelet was mathematically generated to serve as the source signature. A synthetic 2D seismic section was then produced by performing a convolution operation between this wavelet and each vertical trace of the reflection coefficient model. This fundamental step mathematically replicates how a seismic wave interacts with subsurface layers to produce the data recorded by geophones.

For the classification stage, the K-Means clustering algorithm was selected as the analytical engine. A necessary data restructuring was performed, reorienting the seismic data matrix so that each complete seismic trace was treated as an individual sample for the clustering analysis. To objectively determine the most appropriate number of distinct waveform groups, or "clusters," within the data, the Elbow Method was employed. This heuristic technique involves iteratively performing the clustering for a range of cluster counts (from 2 to 10) and plotting the resulting Sum of Squared Errors (SSE) for each iteration. The analysis of this plot revealed a distinct inflection point at k=5, providing a quantitative justification for partitioning the data into five distinct facies.

The K-Means algorithm was then applied, successfully assigning each of the 23 seismic traces to one of the five clusters based on waveform similarity. The results were presented in two key interpretive plots. The first was a seismic facies map, which used a unique color for each cluster to visualize the spatial distribution of the different waveform groups along the 2D line. The second was a series of plots showing the calculated average waveform for each of the five clusters. This allowed for the characterization of the representative seismic signature for each facies, confirming that the algorithm could successfully differentiate the subtle waveform variations across the simple geological model.

The opening exercise constructs a reflection‑coefficient (RC) matrix (Figure 1a), where the RC represents the contrast in acoustic impedance at the interface of two layers. Acoustic impedance, , is defined as the product of velocity and density, and the reflection coefficient is given by RC=(Z2​−Z1​)/(Z2​+Z1​) (Sheriff and Geldart, 1995).

A zero‑phase Ricker wavelet with a dominant frequency of 50 Hz, total length of 50 ms, and 0.1 ms sampling (Figure 1b) is generated from (Enders and Treitel, 2008). The convolution of this wavelet with each column of the RC matrix yields the synthetic seismic section in Figure 1c, whose polarity reversals mirror the imposed impedance trend.

The seismic matrix is first transposed so each trace serves as an individual feature vector, then standardized before being submitted to K‑means clustering. Plotting the sum‑of‑squared errors for k from 2 to 10 (Figure 2a) shows an elbow—beyond five clusters the error reduction becomes negligible—so k = 5 is selected. The mean waveform of each cluster appears in Figure 2b, the same color palette is applied in Figure 2c, and overlaying those colours on the amplitude section (Figure 2d) illustrates how the data‑driven classes capture lateral impedance variations without relying on any prior geological model.

**2.5D Meandering‑Channel Volume**

Building on the 2D proof-of-concept, the second phase introduced a more geologically realistic model: a meandering river channel within a 2.5D volume. This pseudo-3D model was constructed by generating a sequence of 25 individual 2D slices. A predefined RC signature representing a channelized body was embedded within each slice. The lateral position of this channel signature was then programmatically shifted from one slice to the next, governed by a sinusoidal function. This parametric approach effectively sculpted a three-dimensional, sinuous channel system within the model volume.

A higher-resolution 100 Hz Ricker wavelet was convolved with this 3D RC volume to generate a clean synthetic seismic cube. To better emulate real-world acquisition conditions and test the algorithm's robustness, a layer of Gaussian noise was subsequently added to the entire seismic volume. The K-Means clustering algorithm was then applied to this more challenging, noisy dataset.

The classification successfully identified four distinct and geologically meaningful seismic facies. The principal output was a 2D classification map that clearly and accurately delineated the sinuous path of the channel system, demonstrating the algorithm's efficacy in identifying coherent geological features even when they are partially obscured by random noise. The interpretation was further supported by a three-panel diagnostic plot for a selected inline. This powerful visualization provided a side-by-side comparison of the original noisy seismic data, the resulting cluster map, and a transparent overlay of the two. This composite view offered an intuitive and direct confirmation of the strong correlation between the underlying seismic waveform character and the machine-generated facies classification, validating the method's utility on more complex and realistic data.

To emulate real stratigraphy, the single RC row is extruded into a three‑dimensional array of 25 slices. A sinusoidal lateral shift (±4 trace positions per slice) migrates the base RC pattern horizontally, producing a realistic channel belt. Forward‑ and reverse‑shifted sequences are concatenated so the channel sweeps back and forth, after which the volume is trimmed or padded to exactly 25 slices.

A 50 Hz Ricker wavelet sampled at 2 ms is convolved through the 2.5 volume; band‑limited Gaussian noise (σ = 0.1) is added to mimic acquisition artefacts. Figure 3a contrasts the raw amplitude section with its quadric‑interpolated equivalent, while Figure 3b shows a time‑slice through the channel.

Each trace is again flattened into a row vector and clustered. The elbow test on the noisy data prefers k=4 (Figure 4a). Average waveforms for the four clusters are illustrated in Figure 4b, and their spatial distribution is mapped (Figure 4c). Re‑tiling the 1‑D class strip over the amplitude section (Figures 4d–4e) demonstrates how the algorithm isolates channel versus over‑bank facies despite added noise.

**3D Field Data (F3 Cube)**

The final and most critical phase of this study demonstrated the workflow's application to a real-world 3D seismic dataset from the field. The process began with the ingestion and integration of a 1.08 GB seismic data cube in the industry-standard SEGY format and a corresponding picked horizon, representing the top of the Hugin formation, from a standard tabular data file.

A rigorous Quality Control (QC) process was then executed to validate and align these two datasets. This involved creating a continuous structural surface from the discrete horizon data points and visualizing it as a map, which was then overlaid with the seismic survey's boundaries to confirm complete spatial coverage.

The discrete horizon points were then intelligently interpolated to create a continuous surface that conformed precisely to the seismic survey's inline and crossline grid. From this newly mapped surface, a horizon amplitude attribute was extracted and visualized as a map to highlight any amplitude anomalies directly associated with the event. Finally, the mapped horizon was rendered on a representative vertical seismic inline to visually confirm its correct temporal position and structural integrity.

With the data validated, the core analysis proceeded. A vertical analysis window of 11 samples was defined, centered on the picked horizon and including 5 samples above and 5 below. A systematic extraction process was then initiated, capturing the waveform snippet within this window for every trace location across the entire 3D survey. This yielded a comprehensive dataset composed of thousands of individual waveform vectors.

These vectors were then provided as input to the K-Means clustering algorithm. The Elbow Method was again employed to guide the selection of the number of clusters, with the analysis indicating that eight distinct facies provided an optimal representation of the waveform variability within the target interval.

The algorithm successfully partitioned the waveforms into these eight groups. The final interpretive products were delivered through a suite of advanced visualizations designed to translate the numerical output into geological insight:

1. A high-resolution 2D seismic facies map, which revealed coherent, geologically plausible depositional and structural patterns across the survey area.

2. A detailed plot of the average waveform for each of the eight facies, enabling the assignment of a unique and interpretable seismic character to each group (e.g., a high-amplitude peak, a trough-peak doublet, a low-amplitude complex).

3. A final inline display that overlaid the classification results as colored bars directly onto the raw seismic data, providing an intuitive and powerful link between the original seismic data, the machine-generated classification, and the final geological interpretation.

The open-source F3 cube and a time‑interpreted horizon are loaded, while the second is interpolated across the survey (Figure 5). For every trace, five samples above and five below the horizon are extracted, yielding 11‑sample wavelets that populate a two‑dimensional matrix.

After masking traces where the horizon is undefined, an elbow test up to k=11 suggests k=8 classes (Figure 6a). Mean waveforms for those eight clusters (Figure 6b) reveal systematic differences in polarity, symmetry, and bandwidth. Re‑mapping the labels onto the inline‑crossline grid (Figure 7a) produces a facies‑style attribute that highlights buried channels, sand ridges, and amplitude anomalies. Superimposing colour bars on a single inline (Figure 7b) confirms that changes in wavelet shape correlate with stratal terminations and local amplitude dim‑outs, providing an interpreter‑friendly QC of the classification. **Figure 1** — (a) Reflection‑coefficient matrix with lateral impedance variation confined to a single row. (b) Zero‑phase 50 Hz Ricker wavelet, 50 ms long, 0.1 ms sampling. (c) Synthetic seismic section produced by convolving (a) with (b).

SUMMARY

This paper successfully demonstrates a complete and practical workflow for unsupervised seismic waveform classification, systematically progressing from simple synthetic models to a complex, real-world 3D field dataset. The initial 2D and 2.5D synthetic examples served as a crucial validation phase, confirming that the K-Means clustering algorithm could 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, including data synthesis via convolution, the necessary data restructuring for analysis, and the use of the Elbow Method for objectively determining an optimal number of clusters.

The true power and applicability of this workflow were realized in its application to the 3D real-world seismic dataset. By integrating a picked geological horizon, the analysis was precisely targeted to a specific stratigraphic interval of interest. The methodology demonstrated how to programmatically extract thousands of localized waveform snippets from this interval and subject them to K-Means clustering, resulting in the generation of eight distinct seismic facies. The final interpretive products—a detailed seismic facies map revealing coherent depositional patterns, and plots of the characteristic average waveform for each facies—successfully translate the abstract output of a machine learning algorithm into tangible, geologically meaningful insights. This comprehensive, step-by-step guide serves as a valuable and practical resource for geoscientists, illustrating how unsupervised learning can be effectively integrated into seismic interpretation workflows to foster a more efficient, objective, and insightful characterization of the subsurface.

The true power and applicability of this workflow were realized in its application to the 3D real-world seismic dataset. By integrating a picked geological horizon, the analysis was precisely targeted to a specific stratigraphic interval of interest. The methodology demonstrated how to programmatically extract thousands of localized waveform snippets from this interval and subject them to K-Means clustering, resulting in the generation of eight distinct seismic facies. The final interpretive products—a detailed…..

This tutorial has walked through a three‑step workflow that begins with a hand‑built 2‑D reflection‑coefficient section, scales to a 2.5‑D synthetic meandering‑channel volume, and culminates with horizon‑centred wavelet extraction from the open‑source F3 seismic cube. At each stage a zero‑phase Ricker wavelet is forward‑modelled into the impedance grid, K‑means clustering is applied to the resulting traces, and the elbow criterion is used to determine an objective number of waveform classes. Because the procedure relies only 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, intuitive entry point, alternative unsupervised techniques—hierarchical clustering, self‑organising maps, or Gaussian‑mixture models—can provide additional flexibility by revealing nested or non‑spherical cluster structure. Likewise, tying the horizon‑centred 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 further 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 honours both physics and statistics.

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.

Sheriff, R.E. and Geldart, L.P. [1995]. Exploration Seismology. Cambridge University Press.

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

TABLE OF FIGURES

Figure 1. (a) Reflection‑coefficient matrix with lateral impedance variation confined to a single row. (b) Zero‑phase 50 Hz Ricker wavelet, 50 ms long, 0.1 ms sampling. (c) Synthetic seismic section produced by convolving (a) with (b).

Figure 2. (a) Elbow plot of SSE versus number of clusters for the 2D data; the inflection at k=5 is highlighted. (b) Mean waveform of each cluster rendered against its colored background. (c) Cluster map alone. (d) Seismic section with semi‑transparent cluster overlay.

Figure 3. (a) 2.5D synthetic seismic section (slice 12) shown without interpolation and with quadric interpolation. (b) Map view of seismic waveform classification results, with green dashed line marking slice 12.

Figure 4. (a) Elbow plot for the noisy 2.5‑D data; with optimal k=4. (b) Mean waveforms of the four clusters. (c) Seismic amplitude with Quadric‑interpolated, (d) Seismic waveform classification result section. (d). (e) Amplitude with 40%‑opacity seismic waveform classification result overlay.

Figure 5. Inline 445 of the F3 cube with interpreted horizon in red; samples 300–1250 ms displayed.

Figure 6. (a) SSE elbow test for horizon‑centered wavelets; k=8 chosen. (b) Average wavelets for clusters 1–8, plotted on offset‑sample axis.

Figure 7. (a) Wavelet‑cluster map across the full survey; green dashed line marks A – A’ cross section (inline 445). (b) Inline 445 with colored bars showing waveform cluster membership of horizon‑centered wavelets (±5 samples).