Submission code: XM4Q5X86ZA

See this and other abstracts on our website: https://home.sbgf.org.br/Pages/resumos.php

Unsupervised Classification of Evolutionary Spectral Data Using K-Means Clustering

José Alfonzo (Observatório Nacional ON/MCTI), Guilherme Vanzan, Daniel Franco (Coordenação de Geofísica; Observatório Nacional; Rio de Janeiro; Brazil), Ibsen Gomes, Cosme Neto, André Wiermann

Unsupervised Classification of Evolutionary Spectral Data Using K-Means Clustering

Please, do not insert author names in your submission PDF file

Abstract Summary

Digital signal processing techniques, such as the Fast Fourier Transform (FFT), are extensively employed in geophysics for spectral analysis. However, their application to paleoclimatic time series may present limitations, as these methods typically assume stationarity in variance, an assumption that does not always hold true. To address this issue, Evolutionary Spectral Analysis (ESA) offers a framework for visualizing the temporal evolution of spectral components, thereby revealing quasi-periodic behaviors associated with astronomical cycles, such as Earth's orbital eccentricity. Nonetheless, the presence of outliers, data points with anomalously high amplitudes, can introduce spectral artifacts, including spectral leakage, which may lead to misinterpretations. In this context, the present study proposes a methodology grounded in machine learning to classify ESA maps as a preliminary step toward the detection of such artifacts. An unsupervised learning approach was adopted, employing the k-means clustering algorithm in conjunction with the Silhouette Coefficient (SC) to evaluate cluster quality. The results indicate that a five-cluster configuration provides the most meaningful separation among the spectral patterns analyzed. Although some degree of overlap between clusters was observed, the proposed method proved effective in identifying dominant spectral behaviors, highlighting its potential as a supporting tool for improving the reliability of paleoclimatic interpretations. Future developments may focus on enhancing classification accuracy through the incorporation of additional spectral descriptors or the adoption of semi-supervised learning models, as more labeled data become available.

Introduction

In geophysics, digital signal processing techniques are commonly employed for data analysis and interpretation. The Fast Fourier Transform (FFT), for example, is extensively used to convert signals from the time domain to the frequency domain, enabling the identification of dominant spectral components, a technique widely applied in seismology and magnetometry.

In paleoclimatic time series, however, the amplitude of spectral peaks can vary over time. Conventional power spectral density methods, such as the FFT, Periodogram, Lomb-Scargle, and Multitaper, provide estimates of the average variance density as a function of frequency, considering the entire time series. These methods, however, assume the series is stationary with respect to variance, which is not always the case (Weedon, 2003). When the objective is to identify persistent periodicities associated with specific wavelengths over time, such assumptions may lead to misinterpretations.

The evolutionary spectrogram or Evolutionary Spectral Analysis (ESA) is a graphical technique that enables the visualization of the temporal evolution of spectral signals within a time series (Fazio, 2025; Santos, 2023; Leandro, 2022). This method applies a spectral estimator to a moving (sliding) time window and represents power as a function of frequency (x-axis) and time (y-axis). Within the context of cyclostratigraphy, spectrograms generated via ESA reveal the complex quasi-periodic behavior associated with Earth's orbital eccentricity. For instance, they show that spectral components around the 100-thousand-year band often exhibit intervals of reduced power compared to the more stable 405-thousand-year eccentricity cycle, particularly near 37.5 million years ago (Kodama, 2015).

However, the presence of outliers, data points with amplitudes significantly higher than surrounding values, can introduce spectral artifacts within the windows that include these anomalies. A common artifact is spectral leakage, which may generate spurious frequency components. If not properly identified and mitigated, these artifacts can be misinterpreted as genuine paleoclimatic cycles (Percival and Walden, 1993).

This work presents a methodology for classifying ESA graphs as a preliminary step toward artifact detection. It is expected that this pre-classification will support the creation of a database that enables the use of traditional supervised classification methods by simplifying the labeling of data.

Method

To enable the automatic identification of such artifacts, previously a methodology grounded in artificial intelligence and machine learning was developed, with the objective of flagging spectrogram regions where astronomical signals are deemed unreliable for interpretation. Given the geological nature of the data and the lack of reliable labels across a sufficiently broad dataset, a fully supervised learning approach was not feasible. Instead, the process was initiated using an unsupervised learning strategy, aimed at organizing and characterizing recurring spectral patterns.

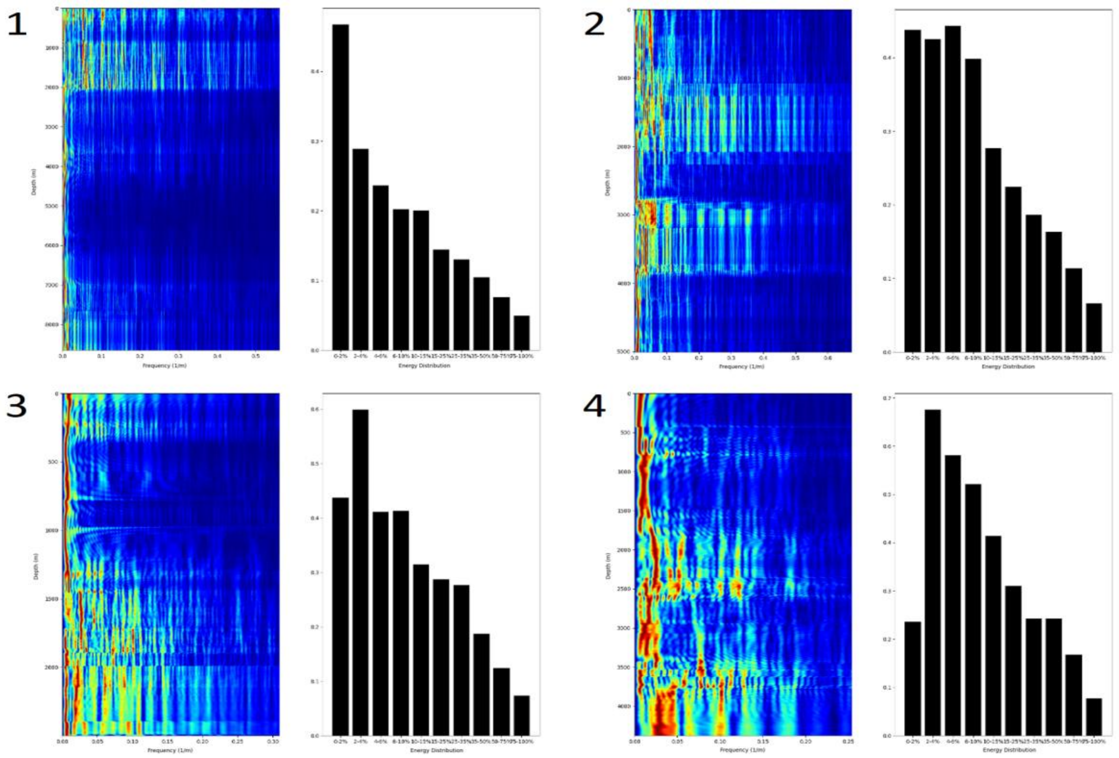

Figure 1: Visual labeling of ESA maps into four groups, considering the visibility and width of the signals across the chart.

The first step involved preprocessing the ESA maps by limiting the maximum frequency to the threshold corresponding to 99% of the total energy in the global spectral analysis. This criterion was adopted to remove high-frequency components that are potentially spurious. Due to the wide variability in the observed spectral patterns, particularly in terms of the width and clarity of spectral features, the k-means clustering algorithm was employed for the classification of the maps (Hartigan, 1979), based on the energy distribution across ten predefined frequency bands: 0–2%, 2–4%, 4–6%, 6–10%, 10–15%, 15–25%, 25–35%, 35–50%, 50–75%, and 75–100%. To guide the interpretation of the classifications performed by k-means, we visually labeled the ESA maps into four groups, considering the visibility and width of the signals throughout the graph (Fig. 1).

The number of clusters (k) is defined prior to each iteration of the k-means algorithm, which returns nine (09) scatter plots of the clustering results, showing the magnitude of the first energy band (0–2%) in relation to the other nine. The minimum number of clusters was set to two (02) and incremented gradually until a decline in the average silhouette score was observed, around seven (07). The optimal number of clusters for this dataset was determined using the Silhouette Coefficient (SC) (Rousseeuw, 1987).

Results

The SC indicated that five clusters (k = 5) provided the most meaningful separation among the types of ESA patterns (Fig. 2). The group with the greatest width and highest visibility (labeled as “4” in Fig. 1) was clearly separated from the rest. For both four and five clusters, group “1” was also well separated. However, groups 3 and 4 showed significant overlap, lacking clearly distinguishable features that would allow for effective separation.

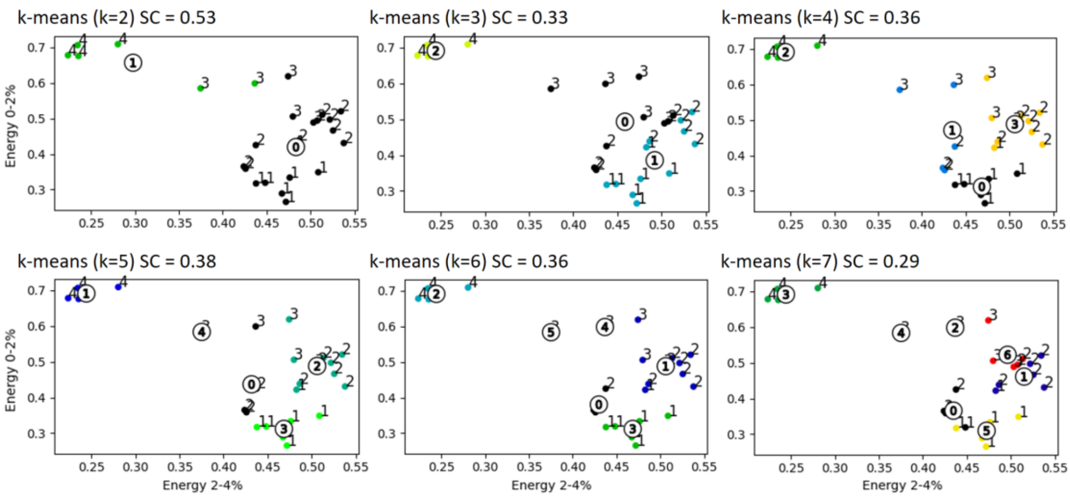

Figure 2: Result of data clustering considering two clusters (k = 2) to seven clusters (k = 7) each accompanied by the corresponding silhouette coefficient (SC) value.

Conclusions

The application of unsupervised machine learning, specifically k-means clustering, to Evolutionary Spectral Analysis (ESA) maps represents a promising approach for the preliminary classification of graphs into groups with similar structures. The use of the Silhouette Coefficient (SC) ensured the selection of an optimal number of clusters, enhancing the interpretability and robustness of the results. Although some overlap remains between certain spectral groups, the methodology successfully identified and categorized dominant spectral behaviors, demonstrating its potential as a support tool for improving the reliability of interpretations. Future work may refine this classification by incorporating additional spectral descriptors or transitioning to semi-supervised models as more labeled data become available.

References

Hartigan, J. A., Wong, M. A., 1979, “A K-Means Clustering Algorithm", Journal of the Royal Statistical Society Series B Methodological, v. 28, n. 1, pp. 100-108.

Weedon, G. P. (2003). Time-series analysis and cyclostratigraphy: examining stratigraphic records of environmental cycles. Cambridge University Press. DOI: 10.1017/CBO9780511535482.

Kodama, K. P., & Hinnov, L. A. (2014). Rock magnetic cyclostratigraphy (Vol. 5). John Wiley & Sons.

Percival, D.B. and Walden, A.T. (1993). Spectral Analysis for Physical Applications. Multitaper and Conventional Univariate Techniques. Cambridge University Press, Cambridge, pp. 1–583.

Rousseeuw, P. J., 1987, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", Journal of computational and applied mathematics, v. 20, pp. 53-65.

Leandro, C.G., Savian, J.F., Kochhann, M.V.L., D. R. Franco, R. Coccioni, F. Frontalini, S. Gardin, L. Jovane, M. Figueiredo, L. R. Tedeschi, L. Janikian, R. P. Almeida & R. I. F. Trindade. Astronomical tuning of the Aptian stage and its implications for age recalibrations and paleoclimatic events. Nat Commun 13, 2941 (2022).

Santos N. B., Rodrigues C. C., Franco D. R., Li M., Fernandes M. A., Candido M., Rocha R. M., Santos T. P., Gasperi A., Fragoso D. G. C., Fazio G., Rodrigues A. N. G; “Integration of seismic stratigraphy and cyclostratigraphy for high resolution chronostratigraphic correlation: The Albacora Field, Campos Basin, Brazil”, Marine and Petroleum Geology, Vol 158, Part B, 2023, 106541.

Fazio G., Yokoyama E., Franco D. R., Rocha R. M., Santos N. B., Cândido M., Leandro C. G., Walde D. H. G.; “Influence of orbital eccentricity on the depositional dynamics of iron formations: Insights from the neoproterozoic Banda Alta Formation (Urucum District, Brazil)” Palaeogeography, Palaeoclimatology, Palaeoecology, Vol 661, 2025, 112715.