Supplement of Nat. Hazards Earth Syst. Sci., 26, 2227–2247, 2026
https://doi.org/10.5194/nhess-26-2227-2026-supplement

Natural Hazards and Earth System Sciences
Open Access
EGU

Supplement of

Seismo-acoustic and GNSS observations of a record-breaking Black Sea storm: repurposing geophysical sensors for environmental monitoring

Laura Petrescu et al.

Correspondence to: Laura Petrescu (laura.petrescu@infp.ro)

S1. Introduction

This supplementary material details the mathematical framework and diagnostic metrics used to categorize single-station infrasound data from the AGIR sensor. To ensure a robust unsupervised classification, we utilized a combination of time-frequency feature extraction and K-Means clustering.

S2. Feature Selection and Mathematical Framework

The input vector for the clustering algorithm consists of five primary signal descriptors: Spectral Flux, Spectral Rolloff, Spectral Entropy, Power Spectral Variance, and Zero-Crossing Rate (ZCR). These features were selected to capture both the harmonic content and the transient energy shifts characteristic of volcanic or atmospheric infrasound. All features were Z-score normalized to prevent variables with larger numerical ranges from biasedly weighting the Euclidean distance calculations in the K-Means algorithm.

Let $x_k[n]$ denote the k-th time window of the signal, and $P_k[m]$ the corresponding power spectrum at frequency bin f_m. A small constant ε is introduced where necessary to prevent division by zero.

The spectral centroid represents the center of mass of the power spectrum and characterizes the dominant frequency content of the signal. It is computed as the first moment of the spectrum:

$$C_k = \frac{\sum_m f_m P_k[m]}{\sum_m P_k[m] + \varepsilon}. \quad (1)$$

Higher values of C_k indicate a shift of spectral energy toward higher frequencies.

Spectral flux measures the magnitude of spectral change between consecutive time windows. It is defined as the ℓ_1 distance between successive power spectra:

$$\Phi_k = \begin{cases} 0, & k = 0, \\ \sum_m |P_k[m] - P_{k-1}[m]|, & k \geq 1. \end{cases} \quad (2)$$

Large values of Φ_k indicate rapid spectral changes, such as those associated with transient acoustic activity.

Spectral rolloff quantifies the frequency below which a specified fraction of the total spectral energy is contained. The rolloff frequency R_k is defined as the smallest frequency satisfying

$$R_k = min\left\{f_{m^*}: \sum_{m \le m^*} P_k[m] \ge 0.85 \sum_m P_k[m]\right\}. \quad (3)$$

This parameter summarizes the upper bound of the dominant energy band of the spectrum.

Spectral entropy measures the complexity or uniformity of the spectral energy distribution. First, the power spectrum is normalized to obtain a probability distribution:

$$\widetilde{P}_k[m] = \frac{P_k[m]+\varepsilon}{\sum_m (P_k[m]+\varepsilon)}. \quad (4)$$

The entropy is then calculated as

$$H_k = -\sum_m \widetilde{P}_k[m] log_2 \widetilde{P}_k[m]. \quad (5)$$

Higher entropy values correspond to broadband or noise-like spectra, whereas lower values indicate concentration of energy in narrow frequency bands.

The zero-crossing count is a simple time-domain measure of signal roughness or oscillatory activity. It is computed from the raw time-series segment as

$$Z_k = \sum_{n=0}^{N-2} 1\{x_k[n] x_k[n+1] < 0\}, \quad (6)$$

where $1\{\cdot\}$ is the indicator function and N is the number of samples in the window. This quantity counts the number of sign changes in the waveform.

S3. Model Optimization and Cluster Validation

To determine the optimal number of clusters (k), we performed a sensitivity analysis using the elbow method and feature pruning:

- **Figure S1: Elbow Method.** The relationship between the number of clusters and the Within-Cluster Sum of Squares (Inertia) is shown. The "elbow" point, where the rate of decrease in inertia significantly levels off, suggests an optimal balance between model complexity and error reduction.
- **Figure S2: Feature Independence and Pearson Correlation.** To address potential multicollinearity, a Pearson correlation matrix was computed for all extracted features. High-intensity coefficients (values > 0.9) were used as a threshold for feature pruning. This step ensures that the K-Means algorithm is not over-sensitized to redundant physical characteristics of the waveform, resulting in more distinct and geophysically meaningful cluster assignments.

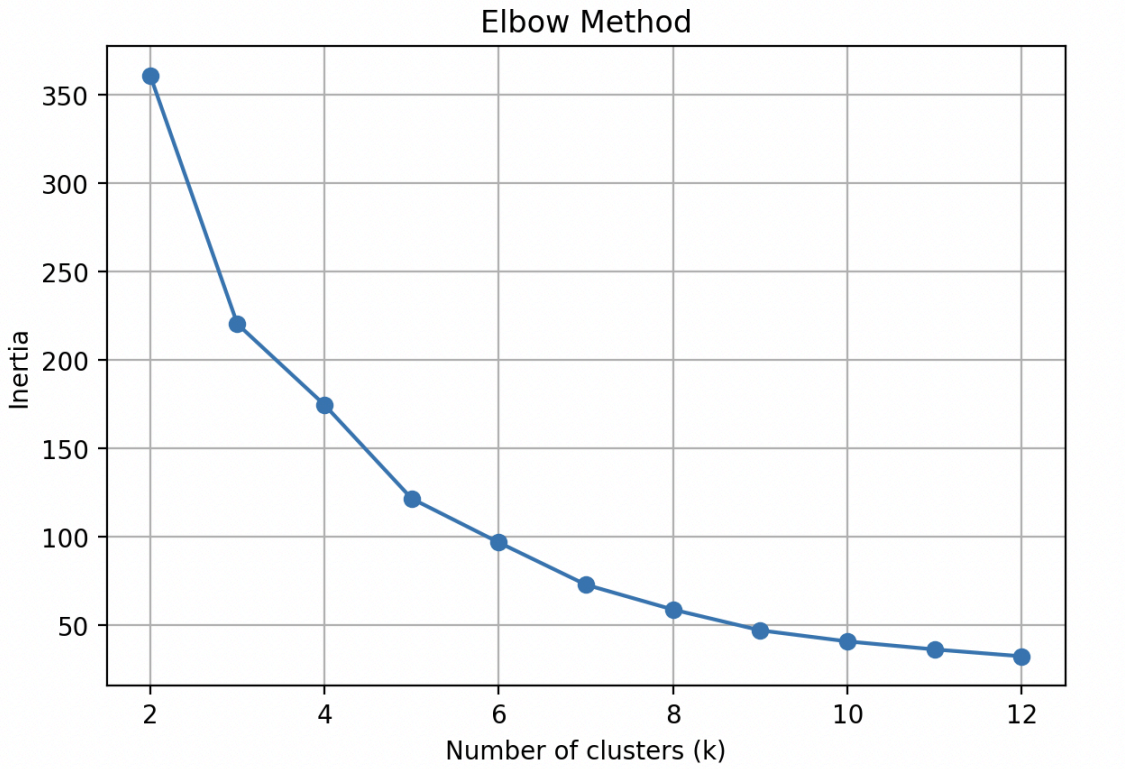

Figure S1. *Determination of the optimal number of clusters (k). The Elbow Method illustrates the relationship between the number of clusters and the total within-cluster sum of squares (Inertia). The elbow point indicates the value of k where additional clusters provide diminishing returns in variance reduction.*

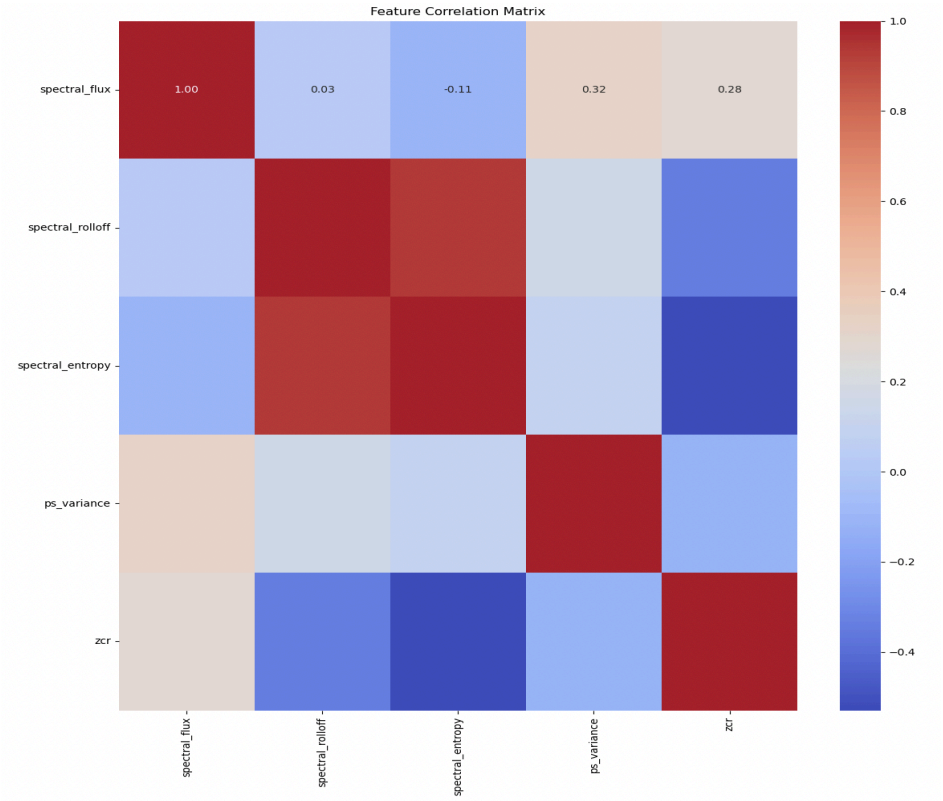

Figure S2. *Pearson Correlation Heatmap of extracted infrasound features. This matrix displays the pairwise correlation coefficients between the time-series features (Spectral Flux, Rolloff, Entropy, Variance, and Zero-Crossing Rate). High-intensity cells (values > 0.90) identify redundant feature pairs. Based on this matrix, a feature pruning step was implemented to remove multicollinearity, ensuring that the K-Means algorithm is not biased toward specific signal characteristics and to improve the numerical stability of the Z-score normalized input space.*