EXPLORING UNSUPERVISED CLUSTERING OF SEISMIC NOISE SOURCES IN URBAN DAS DATA: A METHODOLOGY GUIDE

Antonia Kiel*, Céline Hadziioannou*, Conny Hammer*

* Institute of Geophysics, CEN, University of Hamburg, Germany

antonia.kiel@uni-hamburg.de

To the abstract

Motivation

Since 2021 Distributed Acousic Sensing (DAS) is used to measure the strain rate along a 12 km long optical fiber at the DESY (Deutsches Elektronen-Synchrotron) campus within the *WAVE initiative* [1].

A **large variety of seismic sources with different frequency characteristic** can be observed in the data.

To detect **different types of signals in this large data set, different Machine Learning techniques are compared and a methodology guide is introduced**, recommending which clustering technique to use in different applications.

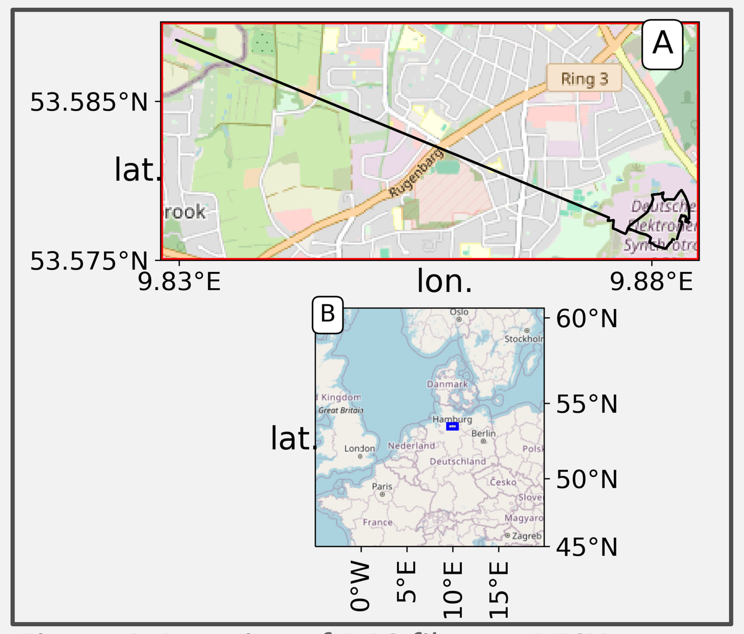


Figure 1: Location of DAS fiber at DESY campus (A) in western Hamburg, Germany (C).

How to use the methodology guide?

1) Which time-frequency represenation to use?

Since the goal is to **detect different seismic signals based on their frequency content**, a time-frequency represenation is needed.

While a spectrogram is the standard way, the continuous wavelet transform (CWT) has a higher time resolution for high frequency signals and a better frequency resolution for lower ones. As a result, the **CWT can represent the frequencies in a more detailed way than a spectrogram**.

The **resolvable frequency range is limited** by the center frequency and bandwidth of the chosen wavelet. For this study, the target frequency of 1 - 80 Hz was sufficiently resolved with a Morlet wavelet with 10 Hz center frequency, so the CWT was preferred.

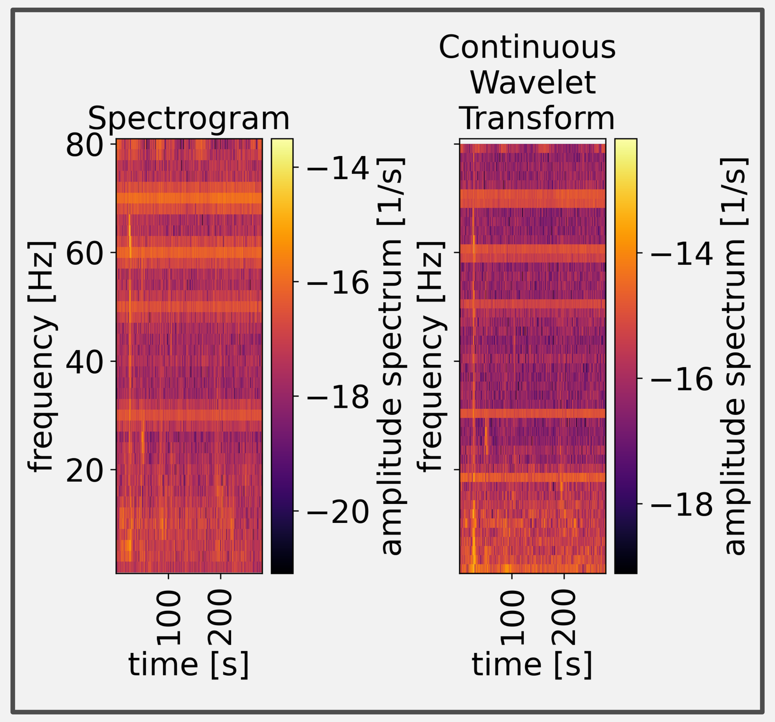


Figure 2: Comparison of spectrogram (left) and continuous wavelet transform (right) for the same data. The signal shows 10 Hz and overtones which are related to transformers located at DESY facilities.

2) Which features to use for clustering?

The time-frequency represenation is an array with the size [time samples x frequency samples]. This can be used as input image to **cluster the data using Deep Embedded Clustering (DEC).** This method reduces the input image to the most important features using a neural network and clustering the latent feature space.

This method can become computationally expensive. Therefore it is useful to **reduce the number of input features**.

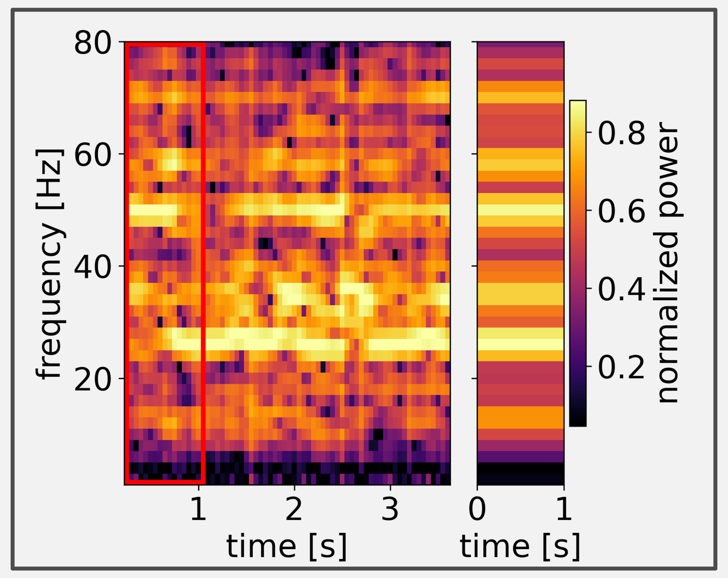
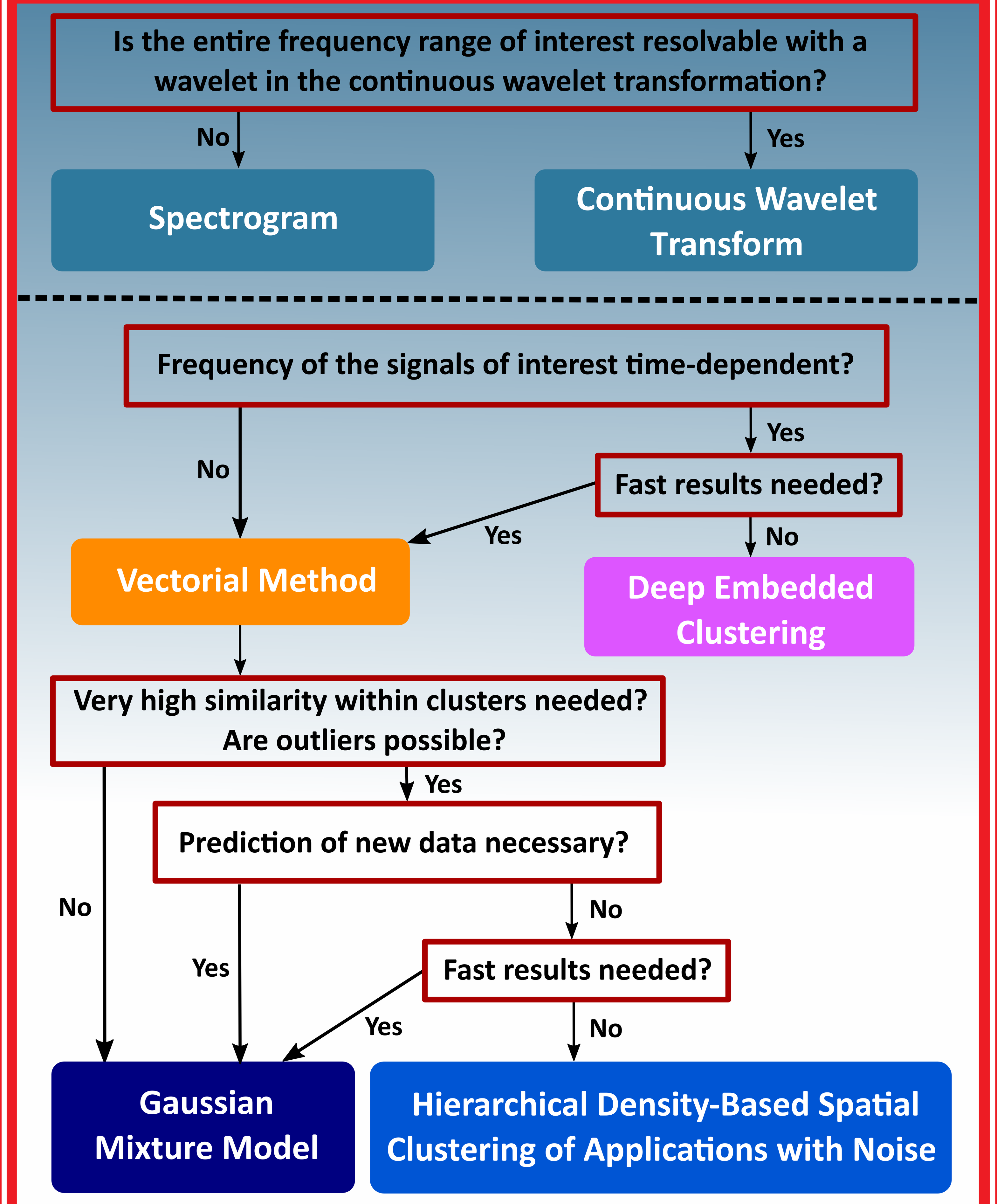


Figure 3: CWT of a few seconds of data (left) and 1-s-average (right) for vectorial method. To improve clustering performance they are normalized.

This is done by **averaging 1 second of data** as introduced by Martin et al., 2018 [2]. This way temporal information is lost but can be justified for signals of consistent frequency. The vectors can afterwards be clustered using standard clustering techniques. In this study the **Gaussian Mixture Model (GMM) and hierarchical density-based spatial clustering of application with noise (HDBSCAN)** performed better than fuzzy-c-means and hierarchical clustering. Therefore only GMM and HDBSCAN are recommended in the guide.

Methodology Guide



Conclusion & Outlook

In this study only the frequency component of data is analysed. High-resolution **spatial features of DAS can be included** in the future by adding time-space represenation to the input. This can e.g. be done by extending the vector to twice its length with the second half being the average of the time-space representation.

The methodology guide can be **applied to many different applications** to cluster data without the need to compare many clustering techniques.

One example is the vectorial method using GMM to monitor seismic source activity at a DAS fiber (e.g. at DESY campus) **in near real-time.**

References

[1] WAVE initative, wave-hamburg.eu

[2] Martin, E., Huot, F., Ma, Y., Cieplicki, R., Cole, S., Karrenbach, M., & Biondi, B. (2018). A seismic shift in scalable acquisition demands new processing: Fiber-optic seismic signal retrieval in urban areas with unsupervised learning for coherent noise removal. IEEE Signal Processing Magazine, 35, 31–40. doi: 10.1109/MSP.2017.2783381

Acknowledgement

We thank the WAVE Team, which includes members of the University of Hamburg, Deutsches Elektronen-Synchrotron (DESY), German GeoForschungZentrum (GFZ) and Helmut Schmidt University, for installing and operating the DAS measurements and discussion about the results.

This work used resources of the Deutsches Klimarechenzentrum (DKRZ) under project ID ug1326.

Example: Deep Embedded Clustering

In this example, the goal is to detect sweeps of active seismic measurements. The vibro truck produced three times four consecutive sweeps with a short break in between.

Since the **frequency of sweeps is time-dependent, DEC is recommended.** However, if fast results are needed, the vectorial method would be preferred.

While the DEC detects only the sweep-related cluster (blue) during excitation, the GMM is dominated by the sweep cluster (red) but significant parts of sweep data are assigned to other clusters. Therefore the **GMM to perform significantly worse on larger data sets with more clusters.**

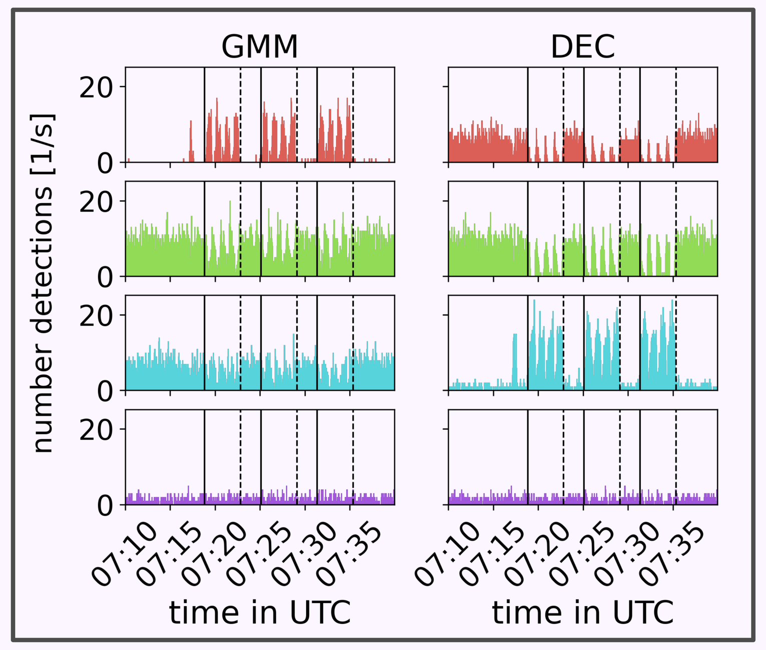


Figure 4: Number of detections of every cluster during active seismic measurements. Horizontal lines show start (solid) and end (dashed) of excitation. GMM (left) is compared to DEC (right) with both methods finding four clusters. For GMM the red cluster correlates with sweeps while for DEC the blue cluster shows seismic sweeps.

Example: Vectorial Method (after Martin et al., 2018)

In this case, the goal is to find seismic sources during two weeks of recording. Most signals like power transformers (figure 2) produce **signals of consistent frequency, so the vectorial method** is used.

Which clustering algorithm to chose is based on the goal of the analysis.

Gaussian Mixture Model

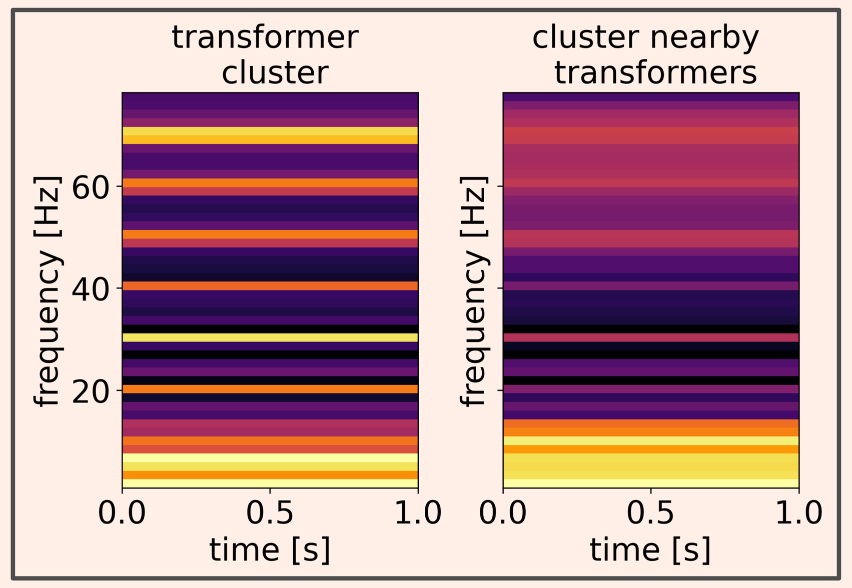


Figure 5: Data example of frequency characteristics of cluster related to transformer signals of 10 Hz and overtones (left). Nearby, another cluster (right) containing attenuated transformer signals is detected by GMM.

Advantages

Every sample is assigned to a cluster.

Relation of **new data** to one of the initially detected clusters **can be predicted.**

Much **faster than HDBSCAN** (here 21 times faster with 1.5 minutes vs. 34 minutes).

→ An example application for GMM is real-time monitoring of active seismic sources to investigate the entrie wavefield at the DAS fiber.

HDBSCAN

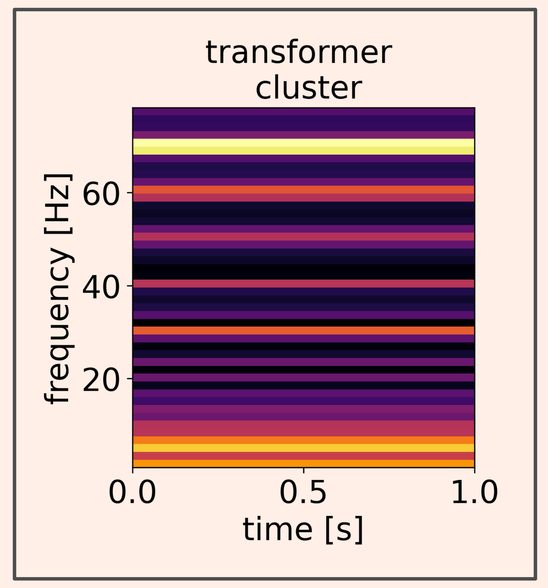


Figure 6: Data example of frequency characteristics of cluster related to transformer signals calculated using HDBSCAN.

Advantages

Much higher **similarity within cluster** of data (93.41 % vs. 96.78 %).

Allows outliers in data set.

→ A potential application is using the detected seismic frequency characteristics to eliminate persistent noise sources for seismic interferometry.