Project report for SCEC award # 25260:
Using both waveform coherence and complexity to cluster microearthquakes

Wenyuan Fan & Peter Shearer
Scripps Institution of Oceanography, UC San Diego

1 Abstract

Earthquake waveforms encode critical information about fault architecture, stress heterogeneity, and fault-zone material properties. Traditional clustering of microearthquakes relies primarily on cross-correlation coefficients, which capture average waveform similarities but do not fully characterize the complexities of entire wavetrains. In this project, we developed a new earthquake clustering procedure that combines waveform coherence (cross-correlation) with an unsupervised machine learning method, the sequencing algorithm (Baron and Ménard, 2021), to cluster microearthquakes based on their full waveform characteristics. The sequencing algorithm orders a set of seismic records by maximizing the similarity between adjacent waveforms (e.g., Carr and Olugboji, 2024), capturing gradual changes in waveform complexity that conventional metrics overlook (Kim et al., 2020; Fang, 2024). This approach clusters events based not only on space and time but also on their entire wavetrains, providing a new means to detail earthquake source properties and fault-zone conditions. We have successfully developed and tested this method and applied it to the 2024 Mw 7.5 Noto Peninsula earthquake sequence in Japan, where a prolonged foreshock sequence offers a unique window into the preparatory processes of a major earthquake. Our results demonstrate that the method can effectively identify seismicity bursts and resolve high-resolution in-situ V_P/V_S ratios, yielding new insights into the temporal evolution of fault-zone stress, fluid conditions, and fracture networks prior to large earthquakes.

2 Achievements

2.1 Method Development

The central goal of this project was to develop a machine-learning-enabled clustering procedure that leverages both waveform coherence and complexity to group microearthquakes. We have accomplished this goal. Our procedure consists of three main steps. First, we align seismic waveforms using cross-correlation to resolve relative time differences between events, similar to back-projection analyses used for imaging large earthquakes. Second, we apply the sequencing algorithm to the aligned waveforms. The sequencer is an unsupervised machine learning method that identifies trends in time series using graph-based metrics, including the Earth Mover distance and the Energy distance, which are sensitive to gradual waveform changes. The algorithm orders the full set of earthquake waveforms to achieve maximum similarity between adjacent records. Third, we apply a minimum spanning tree-based clustering algorithm to the sequenced waveforms to group earthquakes into distinct clusters. This three-step approach successfully clusters events without requiring extensive parameter tuning, and it considers the waveforms of an entire earthquake population rather than evaluating only pairwise similarities.

A key advantage of the method is that the identified clusters yield high-quality differential arrival time measurements for both P- and S-waves, which can be directly used for resolving relative earthquake locations, estimating relative focal mechanisms, and inverting for in-situ V_P/V_S ratios. These capabilities make the clustering procedure a versatile tool that integrates with a range of existing earthquake analysis workflows.

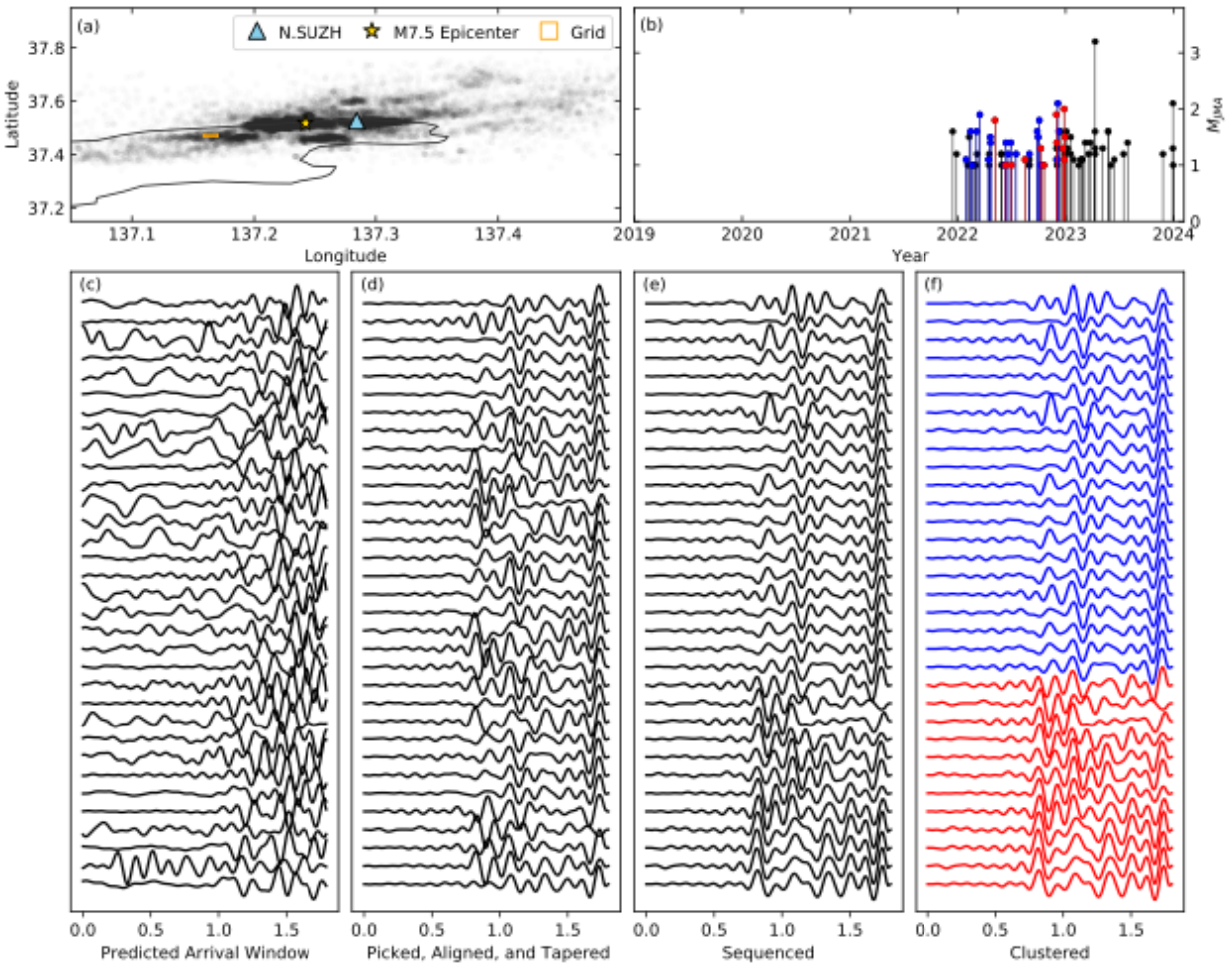

Figure 1: An example of using the sequencer and waveform cross-correlation to cluster earthquakes. (a) Map view of the 2024 Mw 7.5 Noto Peninsula earthquake sequence (Yoshida et al., 2024). The blue triangle indicates the seismic station; the orange square represents a 1 km by 1 km grid for the clustering exercise; the yellow star marks the Noto earthquake hypocenter. (b) Stem plot showing the time and magnitude of microearthquakes within the orange square in (a). Blue and red dots represent two identified clusters using our proposed procedure. (c) S-waves of the earthquakes in the orange square. (d) Aligned S-waves based on waveform cross-correlations. (e) Sequenced waveforms using the aligned waveforms. (f) Two identified clusters based on the waveform sequence. Blue and red traces correspond to the two clusters shown in (b).

2.2 Application to the 2024 Noto Peninsula Earthquake Sequence

We applied our clustering method to the 2024 Mw 7.5 Noto Peninsula earthquake sequence in Japan, which provides an exceptional natural laboratory for testing the procedure. This earthquake was preceded by a four-year foreshock sequence comprising fluid-driven seismic swarms and aseismic slips, offering a rare opportunity to investigate the dynamic preparatory stage and nucleation processes of a major earthquake (Yoshida et al., 2024).

Using seismicity from the Japan Meteorological Agency (JMA) unified earthquake catalog, we employed our new statistical methods to identify seismicity bursts throughout the sequence. Our analyses revealed several important findings. First, the detected seismicity bursts migrate upwards and towards the hypocenters of major $M \geq 5$ foreshocks, and these bursts occur episodically, suggesting pulsed fault-zone processes such

as pore pressure transients and aseismic slip episodes. Second, we resolved high-resolution in-situ V_P/V_S ratios across the fault zone. These measurements are sensitive to Poisson's ratio, fluid pore pressure and saturation, and crack geometry, providing a direct window into fault-zone material properties. We observe anomalous and temporally varying V_P/V_S ratios throughout the fault zone, indicating heterogeneous and evolving fluid and crack conditions. Notably, we find a systematic decrease in the V_P/V_S ratio over the year prior to the mainshock, which is potentially consistent with the organization of fracture networks in advance of the rupture.

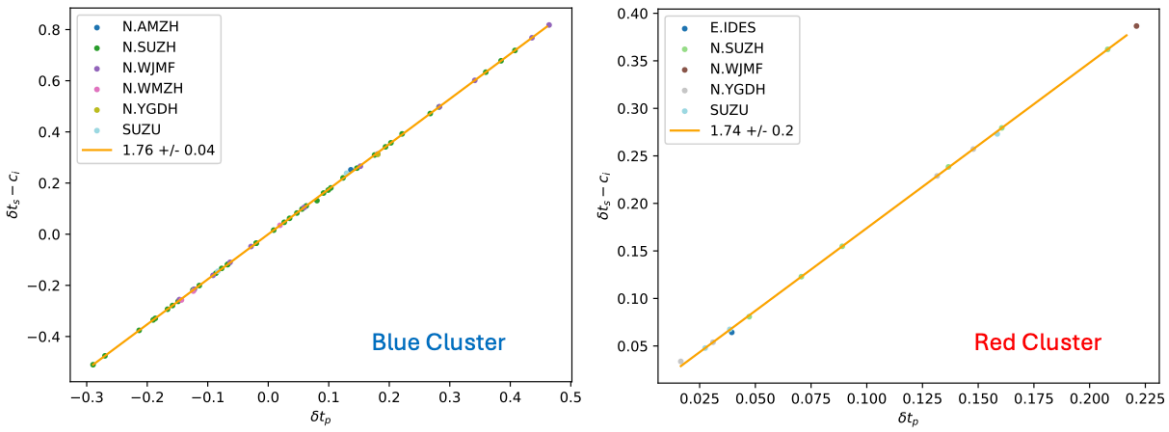

Figure 2: In-situ V_P/V_S estimates of the blue and red clusters in Figure 1. We obtained an estimate of 1.76 for the blue cluster and 1.74 for the red cluster, and their respective uncertainties. We note that the measurements are remarkably clean and consistent, in contrast to typical measurements for such procedures; our clustering procedure yields this set of high-quality measurements.

By comparatively studying the temporal evolution of fault stress and fluid conditions, our work provides new insights into the preparation and nucleation processes of large earthquakes and the physical mechanisms driving seismic swarms.

2.3 Manuscript in Preparation and Student Support

A manuscript presenting the full results of the Noto Peninsula application is currently in preparation. This work constitutes a core component of the Ph.D. thesis of Nicolas DeSalvio, a graduate student supported by this project. The project has provided DeSalvio with training in seismic data analysis, machine learning methods, and fault-zone process interpretation, contributing to his development as an early-career researcher in observational seismology.

3 Conclusions and Outlook

We have successfully developed a new earthquake clustering procedure that combines waveform cross-correlation with the sequencing algorithm, an unsupervised machine learning technique, to cluster earthquakes based on their full waveform characteristics. The method has been tested and applied to the 2024 Mw 7.5 Noto Peninsula earthquake sequence, demonstrating its ability to identify seismicity bursts, resolve high-resolution in-situ V_P/V_S ratios, and illuminate the temporal evolution of fault-zone conditions preceding a major earthquake.

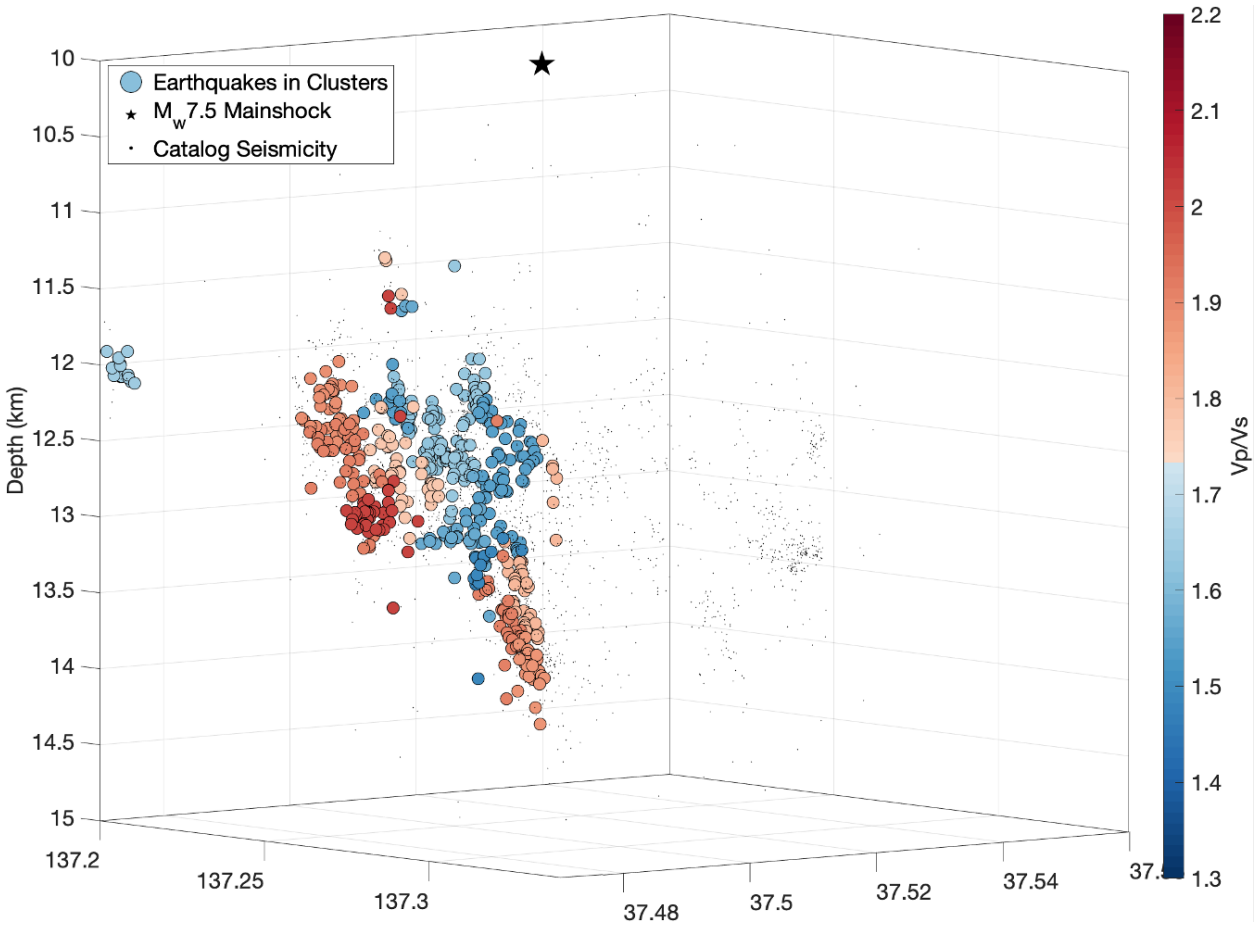

Figure 3: In-situ V_P/V_S estimates near the Mw 7.5 Noto earthquake region, showing high anomalies propagating from deep toward shallower depths. The results imply the presence of possible fluid pathways in the complex fault network.

Looking ahead, several avenues of further development and application are planned. First, we intend to extend the method to incorporate P- and S-waves from multiple stations simultaneously, which will improve the robustness of cluster identification and enable the resolution of more detailed fault-zone structures. Second, we plan to standardize the workflow and release the algorithm as an open-source software package, ensuring broad accessibility and encouraging community-driven applications. Third, we will apply the method to the central San Andreas Fault and the Parkfield segment. These two segments, one dominated by shallow fault creep and the other in a transition zone between seismic and creeping behavior, provide complementary natural laboratories for evaluating the generality of our approach. Finally, the method's integration with earthquake relocation, focal mechanism estimation, and material property imaging positions it as a versatile platform for advancing our understanding of earthquake processes and fault-zone evolution. We anticipate that this tool will contribute to improved seismic hazard assessment and to the broader goals of understanding earthquake preparatory processes.

REFERENCES

Baron, D. and B. Ménard, 2021: Extracting the main trend in a data set: The sequencer algorithm. *The Astrophysical Journal*, **916 (2)**, 91.

Carr, S. A. and T. Olugboji, 2024: A taxonomy of upper-mantle stratification in the us. *Journal of Geophysical Research: Solid Earth*, **129 (5)**, e2024JB028 781.

Fang, H., 2024: Sequencing seismic noise correlations for improving surface wave retrieval and characterizing noise sources. *Seismological Research Letters*, **95 (2A)**, 848–858.

Kim, D., V. Lekić, B. Ménard, D. Baron, and M. Taghizadeh-Popp, 2020: Sequencing seismograms: A panoptic view of scattering in the core-mantle boundary region. *Science*, **368 (6496)**, 1223–1228.

Yoshida, K., R. Takagi, Y. Fukushima, R. Ando, Y. Ohta, and Y. Hiramatsu, 2024: Role of a hidden fault in the early process of the 2024 mw7. 5 noto peninsula earthquake. *Geophysical Research Letters*, **51 (16)**, e2024GL110 993.