Tracing the Central Italy 2016-2017 seismic sequence fault system: Insights from unsupervised Machine Learning and Principal Component Analysis

British Geological Survey

Itahisa González Álvarez[1*], Margarita Segou[2], Brian Baptie[2]

1 British Geological Survey, Keyworth 2 British Geological Survey, Edinburgh

INTRODUCTION

In this work, we investigate a rich deep learning seismic catalogue from the Central Italy 2016-2017 seismic sequence (Tan et al., 2021) with the aim of identifying and tracing active and potentially hazardous faults, as well as studying their distribution and evolution over the duration of the sequence.

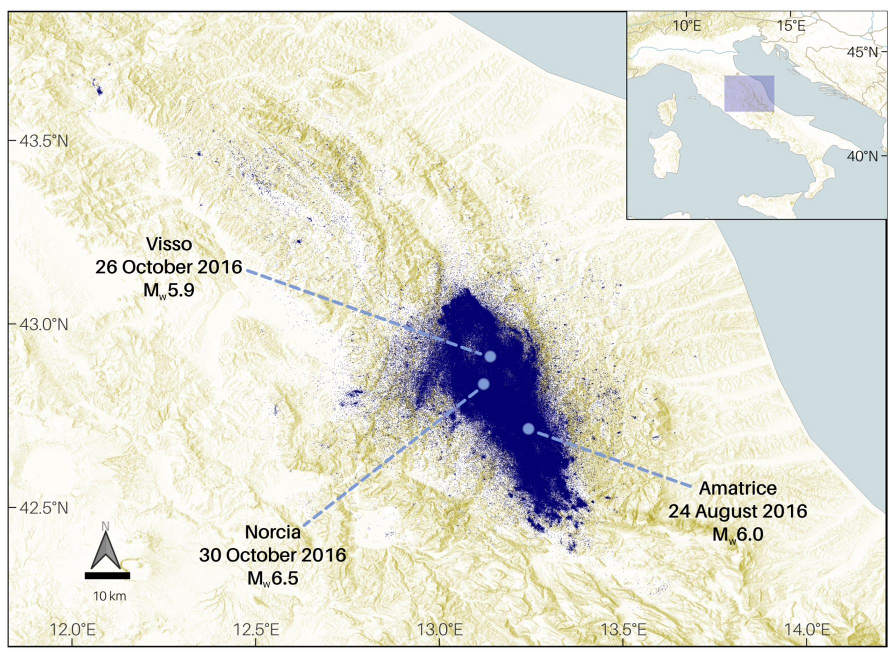
To do this, we tested a variety of unsupervised ML algorithms such as HDBSCAN, DBSCAN, SOM or OPTICS, which we used to design a completely automatic algorithm to identify clustered seismicity. We then combined it with Principal Component Analysis to analyse resulting clusters and relate them to active faults.

Here, we present some of preliminary results from our preferred approach, which highlight the complexity of the fault system, as well as the potential of this method to successfully trace active faults using exclusively seismic catalogue information.

DATA

Tan et al. (2021) analysed 1 year of continuous data from the Central Italy 2016-2017 seismic sequence. They used the deep-neural-network phase picker PhaseNet to analyse waveforms from 139 seismic stations and build an enhanced seismic catalogue with over 900 000 earthquakes with moment magnitudes ranging from 0.5 to 6.2 (of which 72 000 contain focal mechanism information) and a magnitude of completeness of 0.5.

Figure 1. Map view of the area of the 2016-2017 Central Italy seismic sequence. Dark blue dots mark the location of the earthquakes included in the catalogue used in this study, dates ranging from 15 August 2016 to 15 August 2017. Light blue dots show the location of the three largest earthquakes in the sequence.



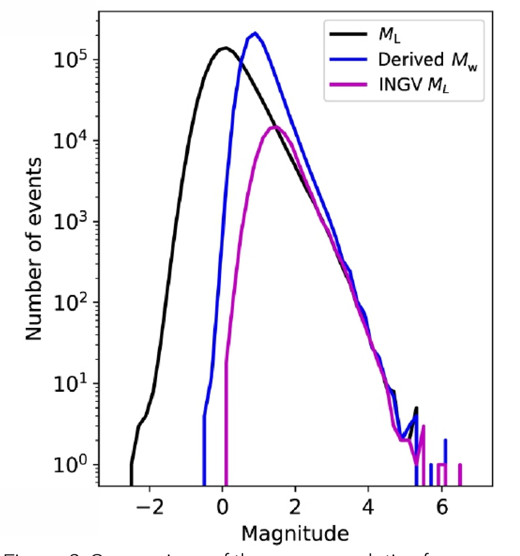


Figure 2. Comparison of the non-cumulative frequency-magnitude distribution of earthquakes in the catalogue used in this study and the one elaborated by the INGV. Modified from Tan et al. (2021).

METHODS

Our work is based on the assumption that seismicity is clustered at, or near, faults. This led us towards density-based clustering methods such as HDBSCAN, DBSCAN, or OPTICS, since earthquakes would tend to cluster tightly around fault planes. Therefore, by applying these methods to our enhanced catalogue, and extracting clusters that represent areas of high density of earthquakes, we can try to relate them to individual active faults.

As the diagram below illustrates, we do this by combining HDBSCAN (McInnes et al., 2017) with our automatic parameter selection algorithm and Principal Component Analysis (PCA). The PCA of individual clusters allows us to define their Principal Plane (PP), which is the surface which explains the most variance of our data (depth and geographical coordinates of the earthquakes in the cluster, in our case) and would be analogous to the fault plane outlined by each cluster. From the PP, we can also obtain an equivalent to the fault's strike and dip that we can compare with the focal mechanisms in our catalogue.

- **Minimum cluster size (mcs):** smallest group of earthquakes that are considered a individual cluster.
- **Minimum number of samples (msam):** number of earthquakes in the neighbourhood for a data point to be considered a core point.
- **Relative Validity (RV):** relative score that allows comparing clustering results from different combinations of hyperparameters. It is an approximation to the Density Based Cluster Validity (DBCV) score.

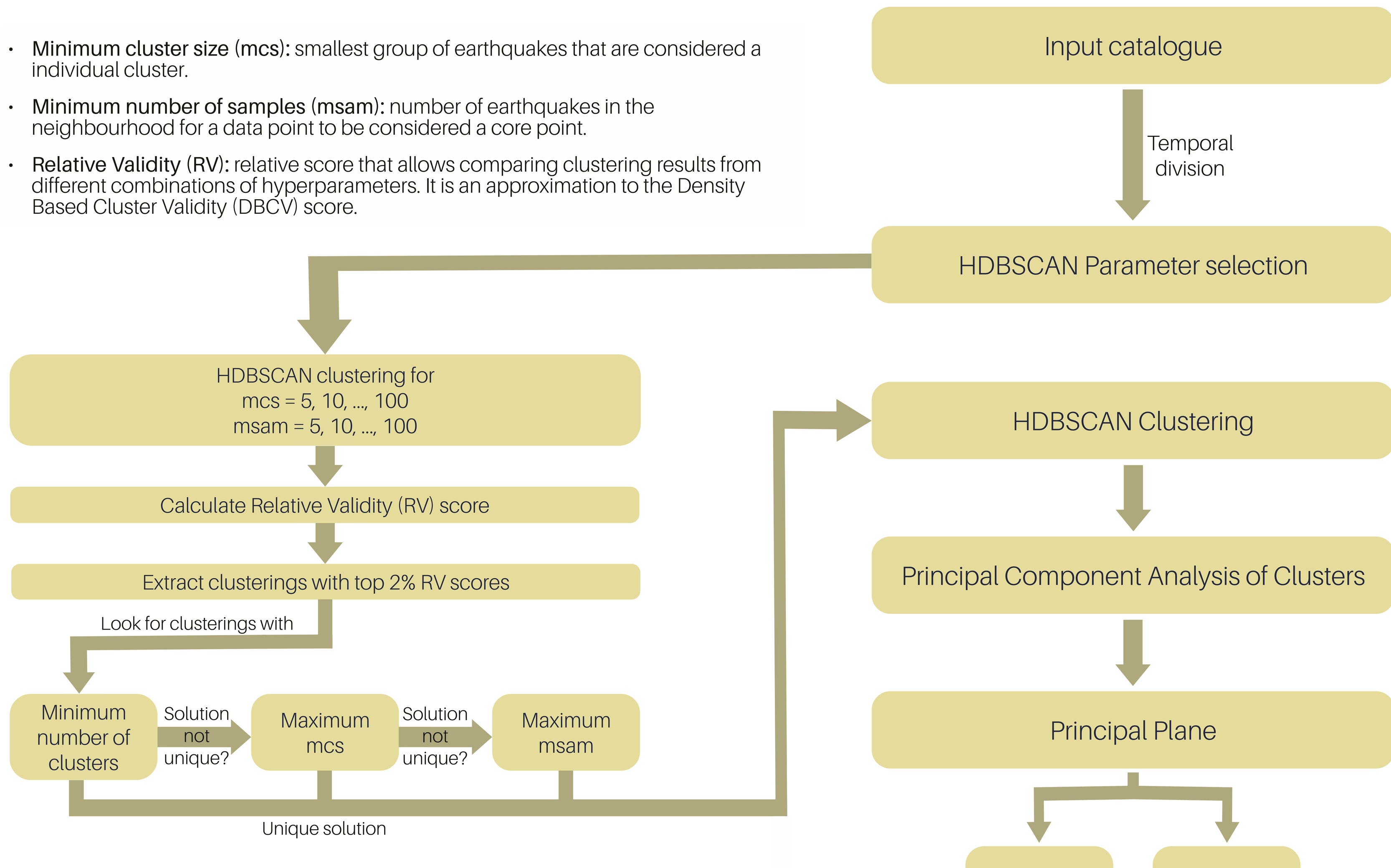


Figure 3. Diagram describing the step-by-step workflow followed in this study.

RESULTS

Our preliminary clustering results of the full, year-long, catalogue, as well as extracted month-, and week-long catalogues, obtained using the algorithm described in Fig. 3, reveal the presence of high-density clusters of earthquakes of varying extent and density within a cloud of diffuse seismicity. PCA then allows us to obtain a Principal Plane for each cluster, from which we can calculate an equivalent to strike and dip that can be compared with the available focal mechanisms for earthquakes within the cluster. This allows us to relate these clusters to individual faults.

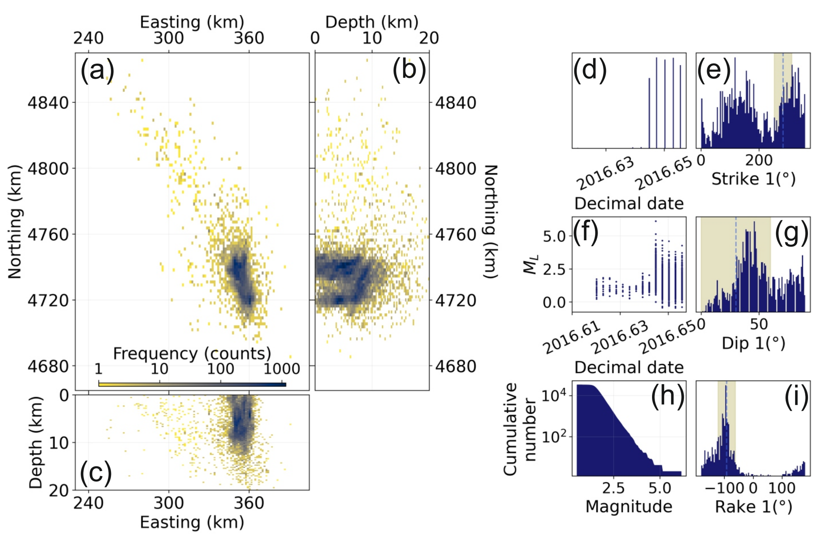
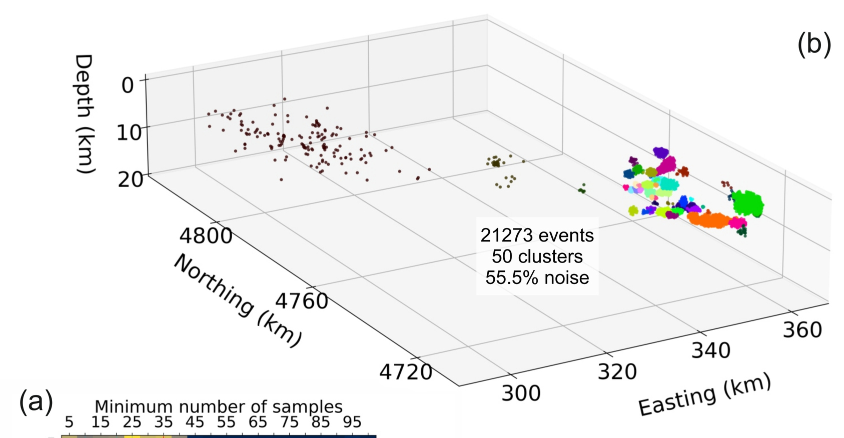


Figure 4. Summary of the main characteristics of catalogue for week 34 of 2016, the second week in the catalogue (panels (a) to (i)). Panels (a) to (c) show the hypocenter density for this week when including unclustered earthquakes from the previous week. Panels (d) to (i) on the right summarize some of the main characteristics of the earthquakes in the week catalogue, including number of earthquakes and magnitudes per day ((d) and (f), respectively), frequency-magnitude distribution (h), as well as available focal mechanism information ((e), (g) and (i)).



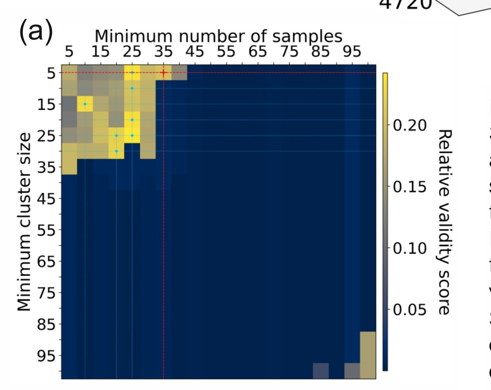


Figure 5. Clustering results for week 34 of 2016, obtained using the algorithm described in Fig. 3. (a) RV score matrix, with blue dots pointing to parameter combinations with an RV within the top 2% of values, and the red square showing the final values used in the clustering. (b) Spatial distribution of clusters obtained for this week catalogue, excluding noise.

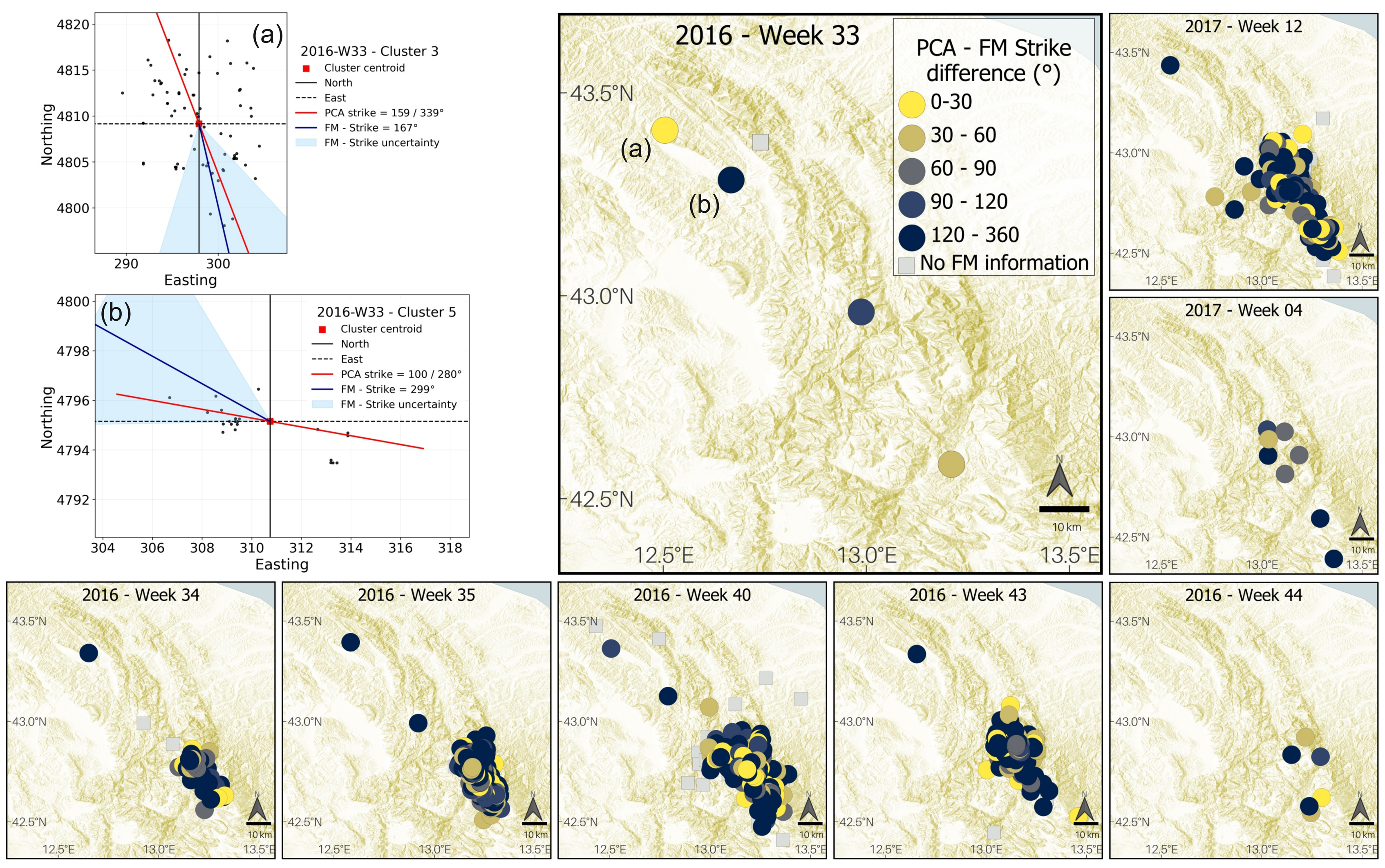
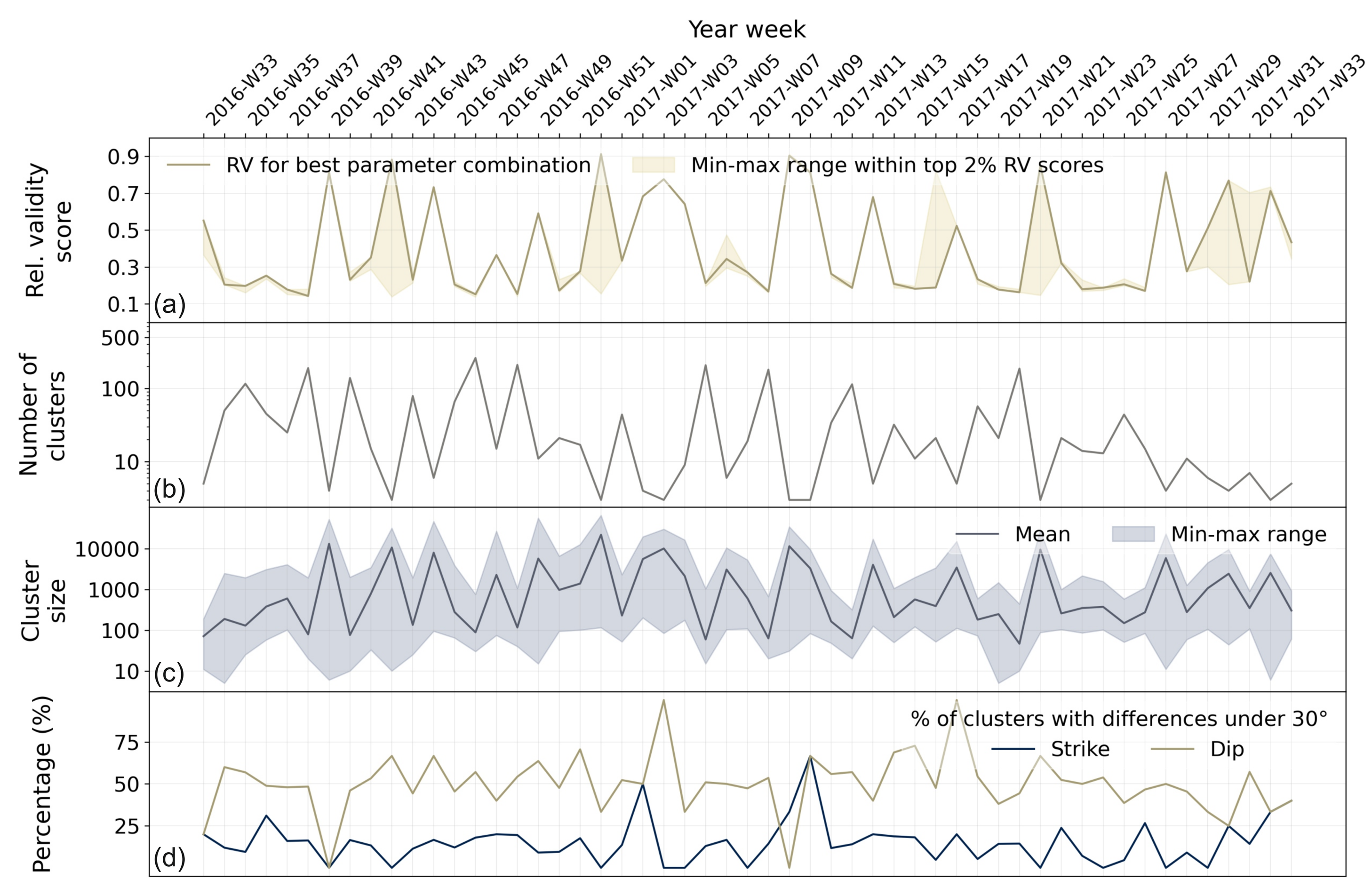


Figure 6. Location of clustering centroids for some of the weeks included in the enhanced catalogue. Colors represent the difference between the strike of the largest earthquake in the cluster with available focal mechanism information and the first strike equivalent obtained from our PCA analysis. Panels (a) and(b) further illustrate these differences by showing the earthquakes contained in clusters 3 and 5 obtained for Week 33 in 2016 (first week in the catalogue), together with their centroids, PC1 vector and strike value and uncertainty from available FMs. It is convenient to note that PCA obtained strike values are affected by an ambiguity, since it is not possible to distinguish between the shortest and longest angles (shown in the legends in panels (a) and (b)) without knowing which side the Principal Plane dips to.

Figure 7. Summary of clustering results obtained for the full catalogue, when split into calendar weeks, showing: (a) RV score for the final parameter combination obtained from the automatic selection algorithm described in Fig. 3, as well as the range of values included within the top 2% of RV scores; (b) total number of clusters; (c) mean cluster size and range of cluster sizes for each week catalogue; (d) percentage of clusters with differences below 30 degrees between strike and dip from available focal mechanism information and equivalents from PCA.



CONCLUSIONS & FUTURE WORK

The preliminary results shown here point to the potential of our method to help us take full advantage of deep learning, enhanced, catalogues by allowing us to trace and study active faults, as well as their evolution over the duration of the seismic sequence, using exclusively seismic catalogue information.

Future improvements will include synthetic tests, resolving the ambiguity in the PCA obtained strike values mentioned in Fig. 6, and continuing to refine our method to improve our ability to trace individual faults.

Corresponding author: itaga@bgs.ac.uk

References:
[1] Tan, Y.J., Waldhauser, F., Ellsworth, W.L., Zhang, M., Zhu, W., Michele, M., Chiaraluce, L., Beroza, G.C. and Segou, M., 2021. Machine-learning-based high-resolution earthquake catalog reveals how complex fault structures were activated during the 2016–2017 Central Italy sequence. The Seismic Record, 1(1), pp.11-19.
[2] L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering In: Journal of Open Source Software, The Open Journal, volume 2, number 11. 2017.