

A hypocenter clustering workflow to illuminate segmented fault surfaces

E. Piegari, G. Camanni, M. Mercurio, W. Marzocchi

Dipartimento di Scienze della Terra, dell'Ambiente e delle Risorse, Università degli Studi di Napoli Federico II, Naples, Italy

Key points

- We present an automatic workflow to illuminate segmented fault surfaces within hypocenter distributions through cluster algorithms
- Comparison with earthquake focal mechanisms corroborates the procedure
- A hierarchical order of planar fault segments associated with different types of faulting is derived

Abstract

We propose a workflow for the automatic recognition of segmented fault surfaces through earthquake hypocenter clustering without prior information. Our approach combines density-based clustering algorithms (DBSCAN and OPTICS), and principal component analysis (PCA). Given a spatial distribution of earthquake hypocenters, DBSCAN identifies first-order clusters, representing regions with the highest density of connected seismic events. Within each first-order cluster, OPTICS further identifies nested higher-order clusters, providing information on their number and size. PCA analysis is applied to first- and higher-order clusters to evaluate eigenvalues, allowing discrimination between seismicity associated with planar features and distributed seismicity that remains uncategorized. The identified planes are then geometrically characterized in terms of their location and orientation in the space, length, and height. This automated procedure operates within two spatial scales: the largest scale corresponds to the longest pattern of approximately equally dense earthquake clouds, while the smallest scale relates to earthquake location errors. By applying PCA analysis before and after OPTICS, a planar feature outputted from a first-order cluster can be interpreted as a fault surface while planes outputted after OPTICS can be interpreted as fault segments comprised within the fault surface. The evenness between the orientation of illuminated fault surfaces and fault segments, and that of the nodal planes of earthquake focal mechanisms calculated along the same faults, corroborates this interpretation. Our automated workflow has been successfully applied to earthquake hypocenter distributions from various seismically active areas (Italy, Taiwan, California) associated with faults exhibiting diverse kinematics.

Plain Language Summary

Active faults are associated with ongoing movement and seismic activity. Recognizing them within large clouds of earthquake hypocenters is at the same time challenging and crucial for seismic hazard estimates. Here, we present a new automatic procedure that can illuminate fault surfaces and its constituting segments by exclusively using hypocenter locations and their spatial density. We apply our approach to hypocenter distributions from various seismically active areas (Italy, Taiwan, California). The evenness between the orientation of illuminated fault surfaces and fault segments, and that derived from other data sources, corroborates our workflow. This workflow is showed to be an effective tool to derive unbiased fault geometries. It also offers new perspectives for the study of the relationships between seismic activity patterns and fault segment interactions, as well as seismic forecasting.

1 Introduction

Geological faults only rarely occur as individual surfaces, but in most cases are complex structures comprising multiple fault segments (Wesnousky, 1988; Walsh and Watterson, 1989; Childs et al., 1996, 2009; Marchal et al., 2003; Walsh et al., 2003; Manighetti et al., 2009; Delogkos et al., 2020; Camanni et al., 2019, 2021, 2023; Roche et al., 2021). In seismically active areas, fault segmentation has been shown to have a significant impact on how a co-seismic rupture nucleates and propagates during an earthquake (Wesnousky, 1988, 2006;

Manighetti et al., 2007, 2009; Hu et al., 2016; Nissen et al., 2016; Perrin et al., 2016; Cesca et al., 2017; Li and Liu, 2020). When an earthquake rupture progresses on a segmented fault and reaches a fault segment boundary, it can either slow down and stop propagating or jump to an adjacent segment, depending on their distance (Wesnousky, 1988, 2006; Yikilmaz et al., 2015; Wang et al., 2020). Consequently, a comprehensive knowledge of the segmentation of an active fault is fundamental when assessing seismic hazard of an area (Field et al., 2003; Boncio et al., 2004; Woessner et al., 2015; Chartier et al., 2019; Bello et al., 2022).

In seismically active areas, fault surfaces can be overall illuminated by analyzing earthquake hypocenter spatial distributions (Ouillon et al., 2008; Ouillon & Sornette, 2011; Kaven et al., 2013; Wang et al., 2013; Kamer et al., 2020; Brunsvick et al., 2021; Truttmann et al., 2023). Associating earthquake hypocenters to fault surfaces is one of the most challenging open issues in seismic hazard assessment, especially for areas where auxiliary information from surface geology, borehole data and geophysical imaging is not available. In the last decades, several methods based on clustering algorithms have been proposed. Partitional clustering such as k-means (Ouillon et al., 2008) and Gaussian Mixture models (Ouillon & Sornette, 2011) have been used to group earthquakes in clusters labelled as faults if the smallest principal component eigenvalue of every cluster is less than the earthquake location error. An attempt to introduce uncertainty of the spatial location of hypocenters using k-means clustering has been made by Wang et al. (2013). Kamer et al. (2020) have proposed a sophisticated method for faults reconstruction based on agglomerative hierarchical clustering where faults are modelled by Gaussian kernels, which are merged until the information gain measured in terms of a Bayesian criterion is positive. Consecutive application of spectral clustering (Von Luxburg, 2007) and density-based clustering (Ester et al., 1996) have been used by Brunsvik et al. (2021) to reconstruct the main Paganica fault system activated during L'Aquila 2009 seismic sequence. An open-source toolbox built on density-based clustering has been provided by Petersen et al. (2021) to reconstruct the first-order geometry of faults. A visual analysis approach that uses an algorithm based on singular value decomposition to fit planar surfaces has been developed by Wang et al. (2019). Truttmann et al. (2023) have introduced a method for 3D imaging of faults that combines nearest neighbor learning and principal component analysis with a Monte Carlo-based approach to account for hypocenter relocation uncertainties.

However, most of these studies, regardless the used algorithms, focus on identifying individual fault surfaces within clouds of hypocenters, missing the illumination of their potential segmented nature. In this work, in the attempt of doing this step forward, we propose an automatic workflow based on the combined use of two density-based clustering algorithms, DBSCAN (Ester et al., 1996) and OPTICS (Ankerst et al., 1999), and principal component analysis, PCA, (Pearson, 1901). Using in combination these clustering algorithms allows us to derive not only information on large-scale faults (through DBSCAN), but also, once they are defined, to derive the geometry of the fault segments comprised within it (through OPTICS).

The paper is organized as follows. In the next section, we describe the proposed workflow and provide details on how location, size and orientation of planar surfaces are retrieved. In sec. 3, we apply the workflow to different areas (Italy, Taiwan, California) where faults exhibit diverse

kinematics (extensional, reverse, strike-slip). For each case, a table with the results for first and second order planar surfaces is provided, while the results for the main third-order features are provided in the supporting information. In sec. 4, a comparison is made between the orientation of illuminated fault surfaces and that of the nodal planes of earthquake focal mechanisms calculated along the same faults. Finally, in sec. 5 we discuss the obtained results and outline potential and limitations of the proposed procedure.

2. Methods

The proposed automated workflow is illustrated in Fig. 1. To illuminate fault surfaces in a given earthquake hypocenter catalog, the first step consists in performing a cluster analysis by DBSCAN (Fig. 1). The application of DBSCAN allows us to illuminate first-order clusters with the highest number of density-connected seismic events. This algorithm requires two inputs: ϵ , the neighborhood radius, and Z , the threshold minimum value for the number of points in the ϵ -neighborhood. Based on the values of ϵ and Z , DBSCAN groups together into clusters only density-connected hypocenters, i.e. hypocenters whose neighborhoods contain at least Z hypocenters that are in the ϵ -neighborhood of each other or lying on their boundaries; all the rest of hypocenters is considered noise and discarded from the cluster analysis. By varying ϵ and Z , DBSCAN provides a very large number of different solutions, which can be divided into five classes according to Piegari et al. (2022). We choose a solution in the class named Crossover Region (CR). This class includes the cluster solutions that simultaneously satisfy the two conditions: i) number of noise points $< 60\%$ and ii) number of points belonging to the biggest cluster $< 60\%$. These solutions maximize the number of the largest clusters. This is the reason why we start by considering one of them, as it allow us to illuminate the biggest number of (first-order) clusters.

In the next step, to define whether or not a first-order cluster can be associated to a planar feature, a principal component analysis (PCA) is performed (Fig. 1). PCA is classified as an unsupervised machine learning algorithm for linear dimensionality reduction. It consists in diagonalizing the covariance matrix, i.e. finding a new basis of coordinates aligned with the directions that maximize the data variation from the mean. The three eigenvalues λ_1 , λ_2 , λ_3 correspond to the variance of the data in the new (decorrelated) reference system. Following Ouillon et al. (2008), if the two largest eigenvalues λ_1 and λ_2 are much larger than the third one λ_3 , the cluster geometrically defines a planar feature (Fig. 1). Within this plane (or if the condition $\lambda_3 \ll \lambda_1, \lambda_2$ is not satisfied for the first-order cluster), the algorithm OPTICS is then applied to investigate its internal hierarchy, for the identification of density-based, second-order, nested clusters (Fig. 1). The output of OPTICS is a type of graph called reachability plot in which denser regions appear as valleys whose depth and width give information about the density and the spatial extension of the denser nested regions, respectively. PCA is iteratively applied to second-order clusters corresponding to each valley. For each of such second-order clusters, if the condition $\lambda_3 \ll \lambda_1, \lambda_2$ is satisfied, a planar feature is built. OPTICS is then further applied for the identification of third-order planar features within second-order ones, etc., up

to the condition for planar geometry is not satisfied or the eigenvalues are smaller than location errors.

For characterizing the geometry of a recognized planar feature, the two eigenvectors \mathbf{u}_1 and \mathbf{u}_2 , corresponding to the eigenvalues λ_1 and λ_2 (i.e., the two principal components, determined by PCA) are used. The plane defining equation can be written in terms of the orthonormal eigenvector $\mathbf{u}_3 = \mathbf{n}$ related to the third eigenvalues λ_3 as:

$$\mathbf{n} \cdot \mathbf{x} - \mathbf{n} \cdot \boldsymbol{\mu}_C = 0 ,$$

where \mathbf{x} is a point that lie on the plane and $\boldsymbol{\mu}_C$ is the cluster barycenter. The third eigenvector \mathbf{n} can be also used to compute the strike (azimuth of an horizontal line lying on the plane) and dip (angle of maximum inclination of the plane) angles of the plane as follows (Quinn and Ehlmann, 2019):

$$strike = \tan^{-1} \frac{n_1}{n_2} - \frac{\pi}{2}$$

$$dip = \cos^{-1} \frac{n_3}{\|\mathbf{n}\|}$$

where n_1 , n_2 and n_3 are the components of the vector \mathbf{n} . The dimensions (length and height) of the planes depend on data variability along the directions \mathbf{u}_1 and \mathbf{u}_2 and thus can be expressed as functions of λ_1 and λ_2 . According to Ouillon et al. (2008), for sake of simplicity we assume that hypocenter locations are uniformly distributed over the planes. If x is a 1D continuous random variable whose values are uniformly distributed on the interval $[a, b]$, its variance is given by:

$$Var(x) = \int_a^b (x - \mu)^2 \frac{1}{b - a} dx = \frac{(b - a)^2}{12}$$

where $\mu = (a + b)/2$. Generalizing this formula to a 2D case, it follows that an approximation for the length L and the height H of the fault planes can be obtained by the equations: $L = \sqrt{12\lambda_1}$, $H = \sqrt{12\lambda_2}$.

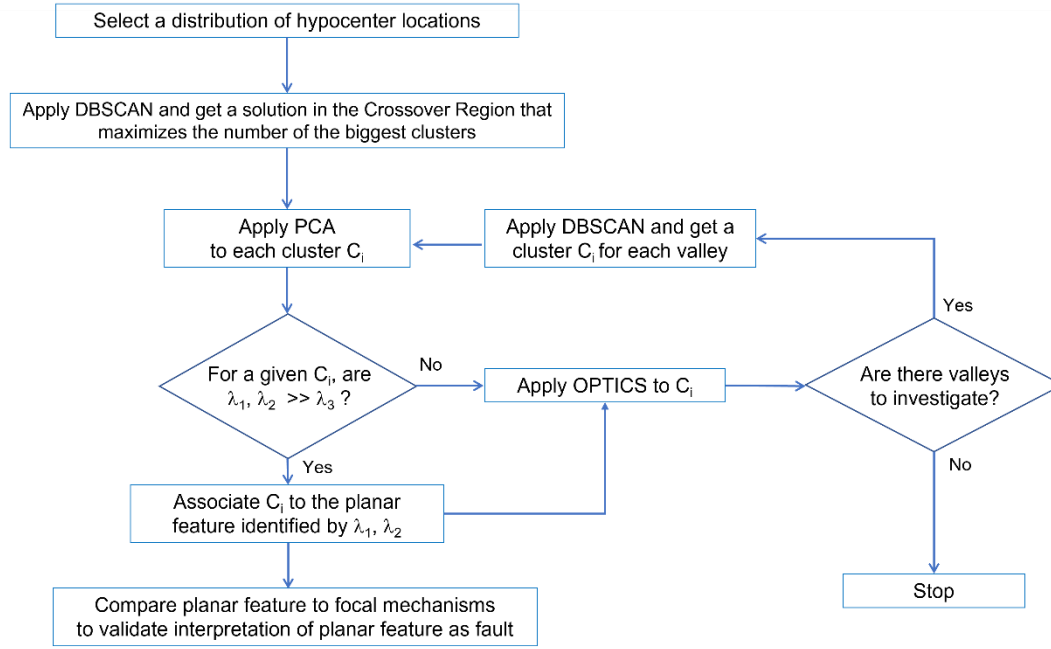


Fig. 1 Flow diagram of the workflow.

3. Workflow applied to case studies

Case 1: clouds of earthquake hypocenters associated with extensional faults (Italy)

As an example of application to a normal fault system, we test the developed workflow by illuminating planar surfaces within the hypocenter distribution of the 2009 L'Aquila (Italy) seismic sequence (Chiaraluce et al., 2012, Lavecchia et al., 2012, Brunsvick et al., 2021). We use the high-resolution earthquake catalog by Valoroso et al. (2013, 2020), which consists of 64051 events spanning from 2009-01-07 to 2009-12-20.

In Fig. 2a, it is shown a DBSCAN cluster solution in the CR corresponding to values of the input parameters $\varepsilon = 0.5$ km and $Z = 200$. Note that before applying DBSCAN, we translated the horizontal coordinates to the hypocenter depth range (2.3 – 21.3 km) using the min-max scaling, which allowed us to consider the spatial anisotropy of the distribution (Piegari et al., 2022). The algorithm automatically illuminates six first-order clusters and discards about 12% of events classified as noise (gray points in Fig. 2a). For brevity, we apply PCA only to the two largest first-order clusters (light blue and light points in Fig. 2) related to L'Aquila and Campotosto faults (Brunsvick et al, 2021), respectively. The results of the applications of PCA with the corresponding geometrical parameters for the reconstructed planar features are reported in Table 1, and the planes are illustrated in Fig. 2b. In addition, to look for second-order clusters within C_1 and C_2 first-order ones, OPTICS was applied. Results (Fig. 3) show that several second-order clusters can be built within first-order ones. Note that all of them can be assigned a planar feature except the smallest second-order cluster C_{14} for which the condition $\lambda_3 \ll \lambda_1, \lambda_2$ is not satisfied. Looking at the reachability plots in Figs. 3a and 3b, it is

visible that many of the named valleys include other sub-valleys that can be associated to third-order clusters. The results of the investigation of the sub-clusters of C_{11} and C_{12} are reported in the supplementary material, along with the animated gif of the 3D plots in Figs. 3c-e.

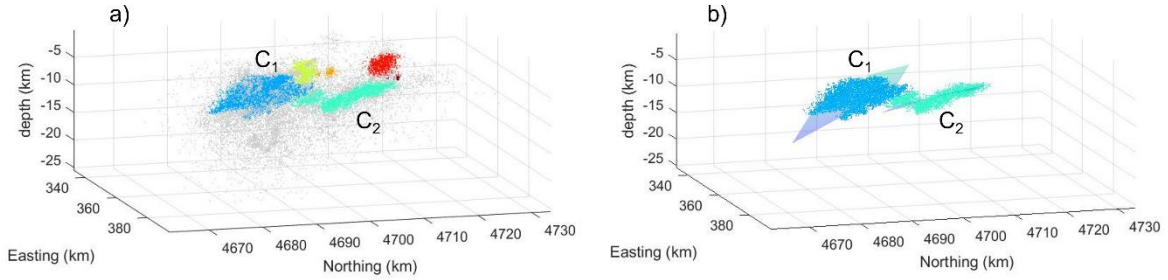
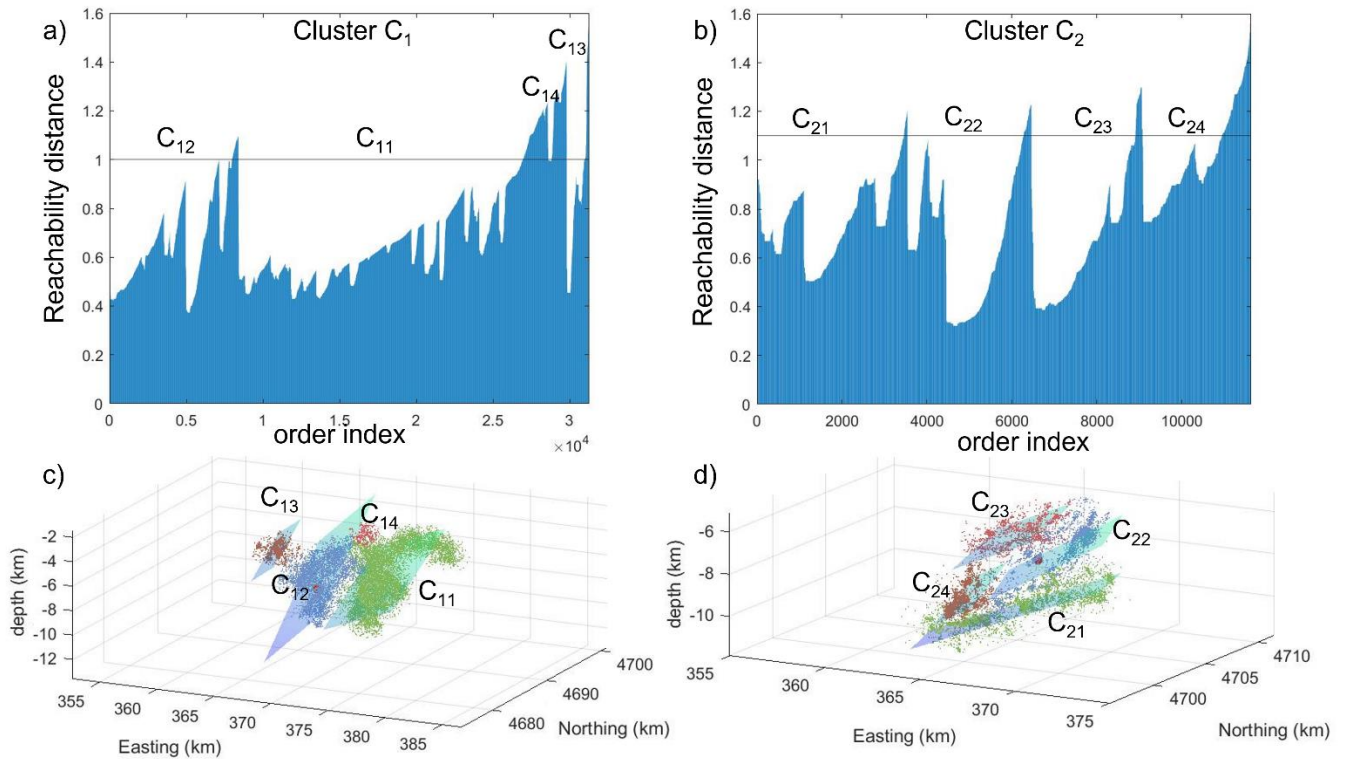


Fig. 2 A workflow application to L'Aquila seismic sequence. a) A DBSCAN solution illuminating six earthquake clusters ($Z = 200$ and $\varepsilon = 0.5$ km). Gray points are earthquakes discarded as noise points. b) First-order planar surfaces retrieved by PCA are shown for the two biggest clusters C_1 and C_2 .



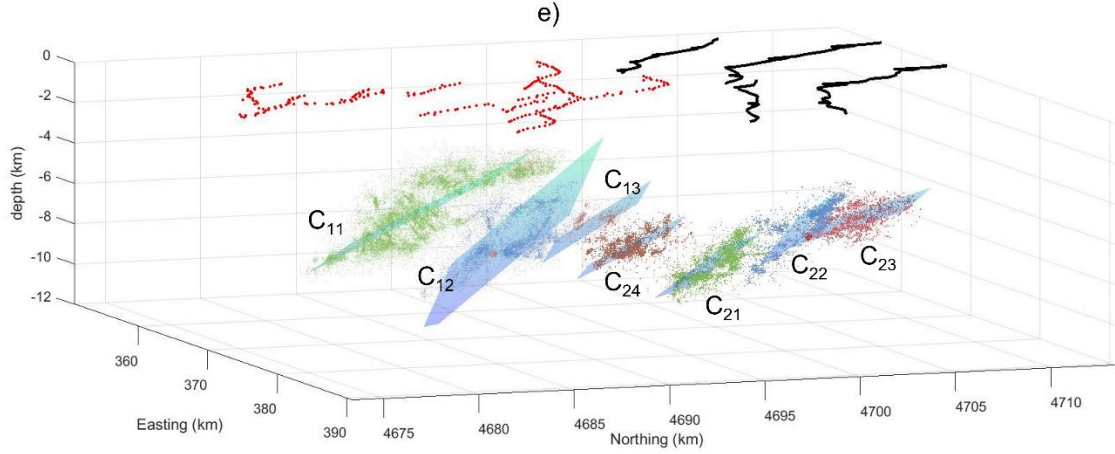


Fig. 3 Reachability plot for clusters C_1 and C_2 in Fig. 2, in a) and b), respectively. DBSCAN solutions for C_1 ($\epsilon = 1$ km and $Z = 200$) and C_2 ($\epsilon = 1$ km and $Z = 200$) with planar surfaces retrieved from PCA in c) and d), respectively. e) A 3D view showing the seven second-order planes retrieved after two iterations of the procedure. The white and red traces at the surface correspond to the faults mapped at the surface in the L'Aquila area (Brunsvik et al., 2021). Earthquakes classified at each iteration as noise points are not shown.

Table 1 Geometrical parameters from PCA analysis applied to first-order clusters C_1 and C_2 in Fig. 2 and related second-order clusters shown in Figs. 3c-e. Longitude, Latitude and Depth are those of the barycenter of the clusters.

Cluster order	Symbol	N	Longitude (°)	Latitude (°)	Depth (km)	L (km)	H (km)	Dip (°)	Strike (°)	λ_3 (km ²)
I (L'Aquila Fault)	C_1	31284	13.451	42.313	-5.7	25.2	8.4	38.9	144.4	0.6880
II (L'Aquila Fault)	C_{11}	18988	13.497	42.285	-4.7	10.3	6.1	29.6	129.5	0.6356
	C_{12}	8446	13.381	42.355	-7.6	10.4	5.4	49.4	140.4	0.1215
	C_{13}	1291	13.291	42.418	-7.9	5.3	2.1	41.7	142.1	0.1340
	C_{14}	328	13.435	42.347	-3.4	-	-	-	-	0.1378
I (Campotosto Fault)	C_2	11634	13.371	42.469	-8.3	18.0	8.0	19.7	151.1	0.6842
II (Campotosto Fault)	C_{21}	3517	13.389	42.449	-9.2	10.4	2.8	25.0	144.5	0.1607
	C_{22}	2871	13.380	42.493	-7.6	6.3	2.5	37.3	137.3	0.2068
	C_{23}	2491	13.383	42.411	-8.2	5.7	2.3	34.7	139.5	0.1482
	C_{24}	2080	13.324	42.532	-7.9	6.3	3.6	28.2	146.8	0.1572

Case 2: clouds of earthquake hypocenters associated with a reverse fault (Taiwan)

Seismicity of the active Taiwan mountain belt is associated with faults with varying kinematics (Kuo Chen et al., 2004, 2007; Wu et al., 2008a, b). We apply the workflow to a spatial distribution of 87679 earthquake hypocenters occurred in the period from 1990-01-01 to 2020-

12-31 in the southeastern sector of the mountain belt (UTM coordinates from 265.99 to 377.5 km Easting and from 2515.9 to 2578.4 km Northing), in an area comprised within the Central and Coastal ranges of the mountain belt (Kuochen et al., 2004; 2007).

Fig. 4a shows a DBSCAN solution in the CR corresponding to $\varepsilon = 1$ km and $Z = 80$. It has been obtained after scaling the horizontal coordinates to the hypocenter depth range (5 – 30 km) to take into account the spatial anisotropy of the distribution. This solution illuminates the biggest, first-order clusters C_1 and C_2 , respectively. We focus our analysis on C_2 as is well-known from the literature that seismicity within it belongs to the Chihshang Fault, a reverse fault that was recently associated with the Mw 6.8 Chengkung earthquake occurred in 2003 (Angelier et al., 2000; Kuochen et al., 2007, 2004; Ching et al., 2007; Mozziconacci et al., 2009). This cluster groups 16244 earthquakes and can be associated to a first-order planar surface (see Fig. 4b) defined by the parameters in Table 2. Fig. 4c shows the reachability plot of C_2 , with the cut line $\varepsilon = 2$ km that identifies the biggest valleys, i.e. two main second-order clusters. Both of them can be associated to planar surfaces (see Fig. 4d), whose parameters are reported in Table 3, and include several subvalleys. They are associated to third-order clusters and are investigated in the supplementary material.

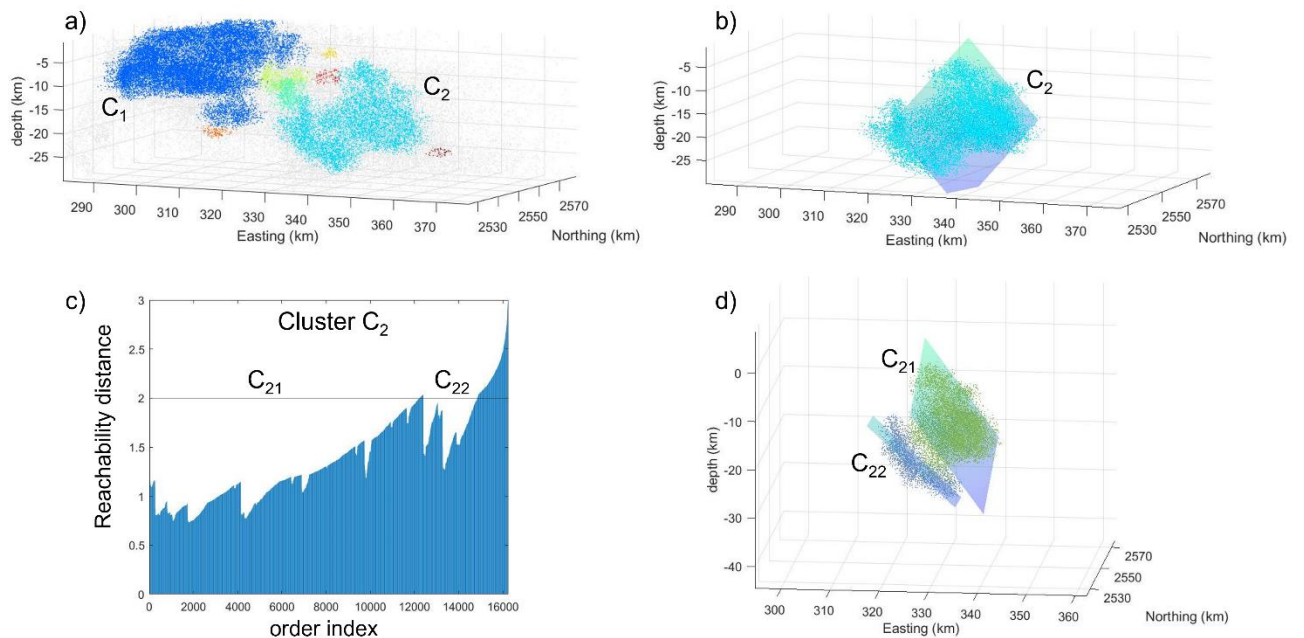


Fig. 4 A workflow application to Taiwan dataset. a) A DBSCAN solution illuminating eight earthquake clusters in a southeastern sector of Taiwan ($Z = 80$ and $\varepsilon = 1$ km). b) First-order planar surface retrieved by PCA is shown for the biggest cluster C_2 . c) Reachability plot of C_2 with the cut line $\varepsilon = 2$ km. d) DBSCAN solution ($\varepsilon = 2$ km and $Z = 80$) with planar surfaces retrieved by PCA. Earthquakes classified as noise points are shown only in the first panel in gray color.

Table 2 Geometrical parameters from PCA analysis applied to first-order cluster C_2 and related second-order clusters shown in Fig. 4b and Fig. 4d. Longitude, Latitude and Depth are those of the barycenter of the clusters.

Cluster order	Symbol	N	Longitude (°)	Latitude (°)	Depth (km)	L (km)	H (km)	Dip (°)	Strike (°)	λ_3 (km ²)
I (Chihshang Fault)	C_2	16244	121.344	23.090	18.3	45.6	16.0	47.6	14.3	3.3626
II (Chihshang Fault)	C_{21}	12227	121.353	23.140	17.9	27.4	15.0	54.3	14.9	1.8778
	C_{22}	2463	121.305	22.894	19.6	18.0	12.5	42.5	2.3	2.5074

Case 3: clouds of earthquake hypocenters associated with a strike-slip fault (California, USA)

We apply the proposed methodology to the most recent swarm occurred in Cahuilla Valley (California, USA), in the area between the Elsinore and San Jacinto strike-slip faults (Ross et al., 2020, Cochran et al., 2023). The swarm seismicity is considered to be controlled by hot fluid circulation, which causes characteristic hypocenter migration patterns with a big number of earthquakes in a relative small volume (Hauksson et al., 2019; Ross et al., 2020; Cochran et al., 2023). We use the high-resolution catalog available at Caltech (2020), which consists of 22700 events spanning from 2016-01-01 to 2019-06-28.

Fig. 5a shows a DBSCAN cluster solution with $\varepsilon = 0.5$ km and $Z = 100$. The algorithm identifies six clusters and discards about 5% of events classified as noise (gray points in Fig. 5a). We apply PCA only to the biggest first-order cluster C_1 (see Fig. 5b) containing nearly all earthquakes (19621 events) and related to a strike-slip fault. Fig. 5c shows the reachability plot of C_1 , with the cut line $\varepsilon = 0.21$ km that identifies several valleys associated to second-order clusters. By applying PCA, we retrieved the geometrical parameters reported in Table 3. The covariance matrices of C_1 and C_{11} have the largest eigenvalues λ_1 and λ_2 of the order of about 1 km, i.e. much larger than relative horizontal and vertical errors (38 m and 87 m, respectively (Ross et al., 2020)). This is the reason why an estimate of the related planar surfaces is attempted even if in both cases the third eigenvalue λ_3 is less than the resolution error (see Table 3). Note that the third eigenvalues of all the remaining second-order clusters are much smaller than the resolution error. Therefore, the related seismicity stays uncategorised.

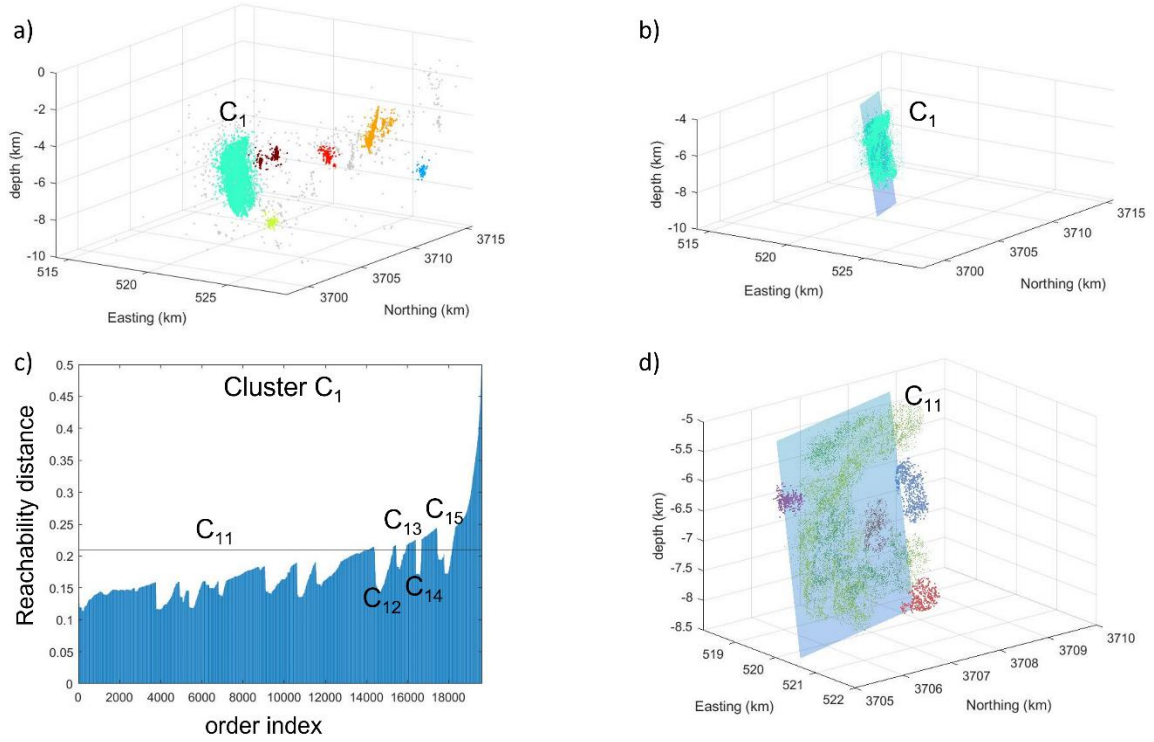


Fig. 5 A workflow application to Cahuilla dataset. a) A DBSCAN solution illuminating six earthquake clusters in Cahuilla Valley ($Z = 100$ and $\epsilon = 0.5$ km). b) First-order cluster C_1 with the planar surface retrieved by PCA. c) Reachability plot of C_1 with the cut line $\epsilon = 0.21$ km. d) DBSCAN solution ($Z = 100$ and $\epsilon = 0.21$ km) with second-order clusters. Earthquakes classified as noise points are shown only in the first panel in gray color.

Table 3 Geometrical parameters from PCA analysis applied to first-order cluster C_1 and related second-order clusters shown in Fig. 5b and Fig. 5d. Longitude, Latitude and Depth are those of the barycenter of the clusters.

Cluster order	Symbol	N	Longitude (°)	Latitude (°)	Depth (km)	L (km)	H (km)	Dip (°)	Strike (°)	λ_3 (km ²)
I (Cahuilla Fault)	C_1	19621	-116.790	33.503	6.8	3.5	3.2	79.9	344	0.015
II (Cahuilla Fault)	C_{11}	14108	-116.789	33.501	6.8	3.4	2.7	80.7	345	0.005
	C_{12}	950	-116.792	33.508	7.0	-	-	-	-	0.001
	C_{13}	737	-116.796	33.518	6.7	-	-	-	-	0.001
	C_{14}	501	-116.789	33.515	8.4	-	-	-	-	3.e-04
	C_{15}	356	-116.786	33.488	6.0	-	-	-	-	0.002

4. Comparison between calculated planar features and fault focal mechanisms

In all areas, orientation of planes retrieved from PCA analysis on both first- and second-order clusters shows a good fit with that derived from nodal planes of focal mechanisms calculated for the same clusters (see Table 4). This is suggesting that the reconstructed planar features for

the first-order clusters can be interpreted as fault surfaces, while those within second-order clusters as fault segments within the embedding fault surface.

Within the same first-order cluster, a planar feature output of PCA analysis before applying OPTICS can be interpreted as a fault surface with a thickness given by the square root of λ_3 , while planes output of PCA analysis after OPTICS can be interpreted as fault segments contained within the fault surface.

Table 4 Comparison between the estimated fault parameters and those retrieved by focal mechanisms.

Case 1: extensional faults (Italy dataset)						
L'Aquila Fault						
Fault angles	DBSCAN+OPTICS +PCA (I-order)	DBSCAN+OPTICS +PCA (II-order)	Brunsvik et al., 2021	Lavecchia et al., 2012	Chairaluce, 2012	Boncio et al., 2010
Dip (°)	~39	~30 – 49	~42.6	~45	~48	
Strike (°)	~N144	~N130 – N142	~N143	~N140	~N137	N130 - N140
Campotosto Fault						
Dip (°)	~20	~25 – 37		~50	~20 ~45 – 50	
Strike (°)	~N151	~N137-147	~N150	N140-150		
Case 2: reverse faults (Taiwan dataset)						
Chihshang fault						
Fault angles	DBSCAN+OPTICS +PCA (I-order)	DBSCAN+OPTICS +PCA (II-order)	Angelier et al., 2001	Kuoehen et al., 2007	Chen et al., 2008	
Dip (°)	~48	~42-54	~39-45	~45	~45-50	
Strike (°)	~N14E	~N2-15E	~N18E	N36.5E		
Case 3: strike-slip faults (California dataset)						
Cahuilla fault						
Fault angles	DBSCAN+OPTICS +PCA (I-order)	DBSCAN+OPTICS +PCA (II-order)	Ross et al., 2020	Cochran et al., 2023		
Dip (°)	~80	~81	~70/80	82		
Strike (°)	344	345	-	343		

5. Discussion

We have presented a new methodology with a dual purpose of illuminating the main fault surfaces within hypocenter clouds and identifying fault segments within them. The results obtained for the L'Aquila seismic sequence show that starting from a DBSCAN solution in the CR allows us to naturally illuminate the main faults (L'Aquila Fault and Campotosto Fault) at the first step as first-order structures. Note that this output is found without the application of any data filter, i.e. we used all data without reduction on magnitudes or spatial extension. The same consideration holds also for the Taiwan and Cahuilla datasets.

Actually, in the case of California dataset since the distribution of earthquakes is characterized by a very high concentration of earthquakes in a relative small volume, our starting point is a

solution not in the CR. This solution is characterized by the biggest cluster C_1 including more than 60% of data and therefore is placed in the region adjacent to the right side of CR in the DBSCAN phase diagram (Piegari et al., 2022). We made this choice in this specific case because considering a solution in the CR (i.e. maximizing the number of large clusters) would have involved subdividing the highest earthquake density region at the first step. This would not have allowed us to identify it as a single first-order structure, as instead it appears evident from the reachability plot. It is indeed characterized by a unique big peak to the right, which identifies as a well-defined single valley all data to its left.

The reachability plots for the other two datasets show a more articulated spatial distribution of earthquakes. In particular, within the first-order Campotosto cluster, four second-order valleys are well evident. They show that the first-order structure is actually not a single surface but it is composed by four main planar segments with different dips that combined together give shape to a listric geometry, in accordance with what derived from other data sources (Chiaraluce et al., 2011). A curved surface also describes the large Chihshang fault (e.g., Kuo-Chen et al., 2007). In this case, the application of the proposed workflow allows us to identify two second-order planar segments with higher and lower dip angles in the northern and southern sectors, respectively. Looking at the reachability plot in Fig. 4c, it is visible that both second-order surfaces are composed by other segments related to the relative sub-valleys. From the analysis of the main third-order clusters, which is reported in the supporting information, a higher dip angle of about 50° is also found for the southern part of large first-order fault structure.

Looking at the reachability plot related to the cluster C_1 associated with L'Aquila fault, third order clusters are easily recognizable as valleys included in the second-order clusters C_{11} and C_{12} (see Fig. 3a). The application of DBSCAN and PCA on such sub-valleys allows us to investigate the geometries of the associated faults with more detail (see supporting information). In particular, the analysis of the valleys of C_{11} shows that the complex morphology of the associated fault can be approximated by two distinct sets of planar segments describing two different types of seismicity: a shallower seismicity concentrated in an almost flat region characterized by very low dip angles (about 10° , see supporting information) and a deeper seismicity related to a fault segment with a vertical extension up to 10 km.

The identification of second and third order planar segments cannot be achieved by a single run of DBSCAN but requires the visual inspection of the reachability plots for the selection of the appropriate ϵ value. This is a limitation of the proposed approach, which at this stage is not fully automatic. However, once the selected order to investigate is chosen, the procedure is very simple and fast, providing parameters and visualization of planar surfaces in a few minutes. The only algorithm of the workflow that is computationally expensive is OPTICS. This is one of the reason why we introduce its use only after the identification of the main first-order largest clusters is made. In this way, the initial amount of data is reduced by a percentage of noise data, which are discarded, and it is conveniently divided in subsets that are more easy to manage. Another limitation of the workflow is related to the use of Euclidean distance as the metric to compute the similarity between data. Since the UTM coordinate system uses map

projections, a relative location error is introduced due to the Earth's curvature. Such an error is usually negligible if the investigated hypocenter distribution extends within the same UTM zone, as in the cases examined in the paper. A further element to improve the workflow is the introduction of uncertainty of attributing earthquakes to clusters. With the goal to keep the approach simple, the output of the presented workflow is an earthquake classification where each hypocenter is strictly assigned to a planar segment or belongs to uncategorized seismicity. Note that a percentage of earthquakes associated to first-order planar surfaces becomes uncategorized when second-order structures are investigated and the same happens when third or higher-order clusters are searched.

The workflow presented in this article has proven to be a valuable tool for roughly illuminating fault surfaces and their segmented nature. However, it is important to note here that this workflow, while providing valuable information for the degree of segmentation of a fault, has certain limitations. For instance, the 3D shape of the tip lines (i.e., boundaries of a fault at which its displacement is null) bounding the automatically retrieved fault surfaces and segments are rectilinear and may differ from their actual shapes. Also, this workflow assumes that the entirety of a fault surface/segments is active, and therefore illuminated by earthquake hypocentres. However, this may not be the case and consequently the automatically retrieved faults may be incomplete surfaces. Therefore, it is essential to exercise caution when using these results. Instead, we recommend using them as proxies to further define fault geometries accurately, a task best carried out by a structural geologist familiar with the 3D geometrical templates of segmented faults, as well as with the geometrical peculiarities of faults of an area.

To end, we outline some perspectives for the future use of the workflow. The study of the spatio-temporal variations of the frequency-magnitude distributions related to the various clusters can provide important implications for seismic hazard forecasting (Herrmann et al., 2022). Furthermore, by splitting the examined hypocenter distribution in different periods, co-seismic interactions between faults of different order could be investigated and compared with the well-known manners of fault interactions derived by analyses of geological (i.e., not currently active) segmented fault surfaces; these may include, for example, exploring modes of soft- and hard-linkage across relay zones between adjacent fault segments associated with displacement transfer processes (see Camanni et al., 2019, and references therein).

6. Conclusions

We propose a methodology exclusively based on unsupervised machine learning algorithms to obtain an unbiased reconstruction of main faults and their segmentation within clouds of earthquake hypocenters. We applied the method to seismic events occurred along faults with diverse kinematics and we were able to estimate fault parameters consistent with those already known by focal mechanisms.

The method allows the reconstruction of faults at multiple scales. The largest one depends on the specific hypocenter distribution and it is bound by the longest pattern of approximately equal dense regions of earthquakes. The lowest one is related to the resolution error.

Iterative loops of DBSCAN, OPTICS and PCA first allow the identification of the largest faults and then illuminate the related segments of decreasing sizes. The retrieved planar surfaces are not randomly oriented but are along specific directions, which indicate faulting directions. The study of their orientations and sizes, also taking into account the temporal dimension, is promising for investigating 3D fault geometries and their spatiotemporal evolution, with important implications on short- and long- term hazard forecasting.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

L'Aquila seismic sequence dataset is available at <https://zenodo.org/record/4036248>; Cahuilla dataset is available at <https://scedc.caltech.edu/data/cahuilla-swarm.html>

The datasets used in this study are stored in the following repository under the CC BY 4.0 license: <https://zenodo.org/record/8210903> (Piegari et al., 2023a). The code to perform the cluster and principal component analyses is implemented in MATLAB environment and makes use of software packages available in the Statistics and Machine Learning Toolbox of MATLAB R2023a. It is permanently stored in the repository <https://zenodo.org/record/8211032> (Piegari et al., 2023b).

References

- Angelier, J., Chu, H. T., Lee, J. C., & Hu, J. C. (2000). Active faulting and earthquake hazard: The case study of the Chihshang fault, Taiwan. *Journal of Geodynamics*, 29(3-5), 151-185. [https://doi.org/10.1016/S0264-3707\(99\)00045-9](https://doi.org/10.1016/S0264-3707(99)00045-9)
- Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2), 49-60. <https://doi.org/10.1145/304181.304187>
- Bello, S., Lavecchia, G., Andrenacci, C., Ercoli, M., Cirillo, D., Carboni, F., Barchi, M. R., & Brozzetti, F. (2022). Complex trans-ridge normal faults controlling large earthquakes. *Scientific Reports*, 12, 10676. <https://doi.org/10.1038/s41598-022-14406-4>
- Boncio, P., Lavecchia, G. & Pace, B. (2004). Defining a model of 3D seismogenic sources for Seismic Hazard Assessment applications: The case of central Apennines (Italy). *Journal of Seismology* 8, 407-425. <https://doi.org/10.1023/B:JOSE.0000038449.78801.05>
- Boncio, P., Pizzi, A., Brozzetti, F., Pomposo, G., Lavecchia, G., Di Naccio, D., & Ferrarini, F. (2010). Coseismic ground deformation of the 6 April 2009 L'Aquila earthquake (central Italy, Mw6. 3). *Geophysical Research Letters*, 37(6). <https://doi.org/10.1029/2010GL042807>

Brunsvik, B., Morra, G., Cambiotti, G., Chiaraluce, L., Di Stefano, R., De Gori, P., & Yuen, D. A. (2021). Three-dimensional paganica fault morphology obtained from hypocenter clustering (L'Aquila 2009 seismic sequence, Central Italy). *Tectonophysics*, 804, 228756. <https://doi.org/10.1016/j.tecto.2021.228756>

Camanni, G., Roche, V., Childs, C., Manzocchi, T., Walsh, J., Conneally, J., Saqab Mudasar M. & Delogkos, E. (2019). The three-dimensional geometry of relay zones within segmented normal faults. *Journal of Structural Geology*, 129, 103895. <https://doi.org/10.1016/j.jsg.2019.103895>

Camanni, G., Vinci, F., Tavani, S., Ferrandino, V., Mazzoli, S., Corradetti, A., Parente, M., & Iannace, A., (2021). Fracture density variations within a reservoir-scale normal fault zone: A case study from shallow-water carbonates of southern Italy. *Journal of Structural Geology*, 151, 104432.

Camanni, G., Childs, C., Delogkos, E., Roche, V., Manzocchi, T., & Walsh, J. (2023). The role of antithetic faults in transferring displacement across contractional relay zones on normal faults. *Journal of Structural Geology*, 168, 104827. <https://doi.org/10.1016/j.jsg.2023.104827>

Caltech (2020). Cahuilla swarm catalog (2016-2019), available at <https://scedc.caltech.edu/data/cahuilla-swarm.html>

Cesca, S., Y. Zhang, V. Mouslopoulou, R. Wang, J. Saul, M. Savage, S. Heimann, S.-K. Kufner, O. Oncken, & Dahm, T. (2017). Complex rupture process of the Mw 7.8, 2016, Kaikoura earthquake, New Zealand, and its aftershock sequence, *Earth and Planetary Science Letters*, 478, 110-120. <https://doi.org/10.1016/j.epsl.2017.08.024>

Chartier, T., Scotti, O., & Lyon-Caen, H. (2019). SHERIFS: open-source code for computing earthquake rates in fault systems and constructing hazard models. *Seismol Res. Lett.* 90, 1678-1688. doi:10.1785/0220180332

Chen, K. H., Nadeau, R. M., & Rau, R. J. (2008). Characteristic repeating earthquakes in an arc-continent collision boundary zone: The Chihshang fault of eastern Taiwan. *Earth and Planetary Science Letters*, 276(3-4), 262-272.

Chiaraluce, L. (2012). Unravelling the complexity of Apenninic extensional fault systems: A review of the 2009 L'Aquila earthquake (Central Apennines, Italy). *Journal of Structural Geology*, 42, 2-18. <https://doi.org/10.1016/j.jsg.2012.06.007>

Childs, C., Nicol, A., Walsh, J. J., & Watterson, J. (1996). Growth of vertically segmented normal faults. *Journal of Structural Geology*, 18(12), 1389-1397. [https://doi.org/10.1016/S0191-8141\(96\)00060-0](https://doi.org/10.1016/S0191-8141(96)00060-0)

Childs, C., Manzocchi, T., Walsh, J. J., Bonson, C. G., Nicol, A., & Schöpfer, M. P. (2009). A geometric model of fault zone and fault rock thickness variations. *Journal of Structural Geology*, 31(2), 117-127. <https://doi.org/10.1016/j.jsg.2008.08.009>

Ching, K. E., Rau, R. J., & Zeng, Y. (2007). Coseismic source model of the 2003 Mw 6.8 Chengkung earthquake, Taiwan, determined from GPS measurements. *Journal of Geophysical Research: Solid Earth*, 112(B6). <https://doi.org/10.1029/2006JB004439>

Cochran, E. S., Page, M. T., van der Elst, N. J., Ross, Z. E., & Trugman, D. T. (2023). Fault Roughness at Seismogenic Depths and Links to Earthquake Behavior. *The Seismic Record*, 3(1), 37-47. Doi: <https://doi.org/10.1785/0320220043>

Delogkos, E., Manzocchi, T., Childs, C., Camanni, G., & Roche, V. (2020). The 3D structure of a normal fault from multiple outcrop observations. *Journal of Structural Geology*, 136, 104009. <https://doi.org/10.1144/SP439.19>

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).

Field, E.H., Jordan, T. H., & Cornell, C. A. (2003). OpenSHA: A Developing Community-modeling Environment for Seismic Hazard Analysis. *Seismological Research Letters*, 74(4), 406-419. <https://doi.org/10.1785/gssrl.74.4.406>

Hauksson, E., Ross, Z. E., & Cochran, E. (2019). Slow-growing and extended-duration seismicity swarms: Reactivating joints or foliations in the Cahuilla Valley Pluton, Central Peninsular Ranges, Southern California. *Journal of Geophysical Research: Solid Earth*, 124(4), 3933-3949. <https://doi.org/10.1029/2019JB017494>

Herrmann, M., Piegari, E., & Marzocchi, W. (2022). Revealing the spatiotemporal complexity of the magnitude distribution and b-value during an earthquake sequence. *Nature Communications*, 13(1), 5087. <https://doi.org/10.1038/s41467-022-32755-6>

Hu, F., Zhang, Z., & Chen, X. (2016). Investigation of earthquake jump distance for strike-slip step overs based on 3-D dynamic rupture simulations in an elastic half-space. *Journal of Geophysical Research: Solid Earth*, 121, 994–1006. <https://doi.org/10.1002/2015JB012696>

Kamer, Y., Ouillon, G., & Sornette, D. (2020). Fault Network Reconstruction using Agglomerative Clustering: Applications to South Californian Seismicity. *Natural Hazards and Earth System Sciences Discussions*, 2020, 1-23. <https://doi.org/10.5194/nhess-20-3611-2020>

Kaven, J. O., & Pollard, D. D. (2013). Geometry of crustal faults: Identification from seismicity and implications for slip and stress transfer models. *Journal of Geophysical Research: Solid Earth*, 118(9), 5058-5070. <https://doi.org/10.1002/jgrb.50356>

Kuoehen, H., Wu, Y. M., Chang, C. H., Hu, J. C., & Chen, W. S. (2004). Relocation of eastern Taiwan earthquakes and tectonic implications. *Terrestrial, Atmospheric and Oceanic Sciences*, 15(4), 647. [https://doi.org/10.3319/TAO.2004.15.4.647\(T\)](https://doi.org/10.3319/TAO.2004.15.4.647(T))

Kuoehen, H., Wu, Y. M., Chen, Y. G., & Chen, R. Y. (2007). 2003 Mw6. 8 Chengkung earthquake and its related seismogenic structures. *Journal of Asian Earth Sciences*, 31(3), 332-339. <https://doi.org/10.1016/j.jseas.2006.07.028>

Lavecchia, G., Ferrarini, F., Brozzetti, F., De Nardis, R., Boncio, P., & Chiaraluce, L. (2012). From surface geology to aftershock analysis: Constraints on the geometry of the L'Aquila 2009 seismogenic fault system. *Italian Journal of Geosciences*, 131(3), 330-347. <https://doi.org/10.3301/IJG.2012.24>

Li, G., & Liu, Y. (2020). Earthquake rupture through a step-over fault system: An exploratory numerical study of the Leech River Fault, southern Vancouver Island. *Journal of Geophysical Research: Solid Earth*, 125, e2020JB020059. <https://doi.org/10.1029/2020JB020059>

Manighetti, I., Campillo, M., Bouley, S., & Cotton, F. (2007). Earthquake scaling, fault segmentation, and structural maturity. *Earth and Planetary Science Letters*, 253 (3-4), 429-438. <https://doi.org/10.1016/j.epsl.2006.11.004>.

Manighetti, I., Zigone, D., Campillo, M., & Cotton, F. (2009). Self-similarity of the largest-scale segmentation of the faults: Implications for earthquake behavior. *Earth and Planetary Science Letters*, 288(3-4), 370-381. <https://doi.org/10.1016/j.epsl.2009.09.040>.

Marchal, D., Guiraud, M., & Rives, T. (2003). Geometric and morphologic evolution of normal fault planes and traces from 2D to 4D data. *Journal of Structural geology*, 25(1), 135-158. [https://doi.org/10.1016/S0191-8141\(02\)00011-1](https://doi.org/10.1016/S0191-8141(02)00011-1)

Mozziconacci, L., Delouis, B., Angelier, J., Hu, J. C., & Huang, B. S. (2009). Slip distribution on a thrust fault at a plate boundary: The 2003 Chengkung earthquake, Taiwan. *Geophysical Journal International*, 177(2), 609-623. <https://doi.org/10.1111/j.1365-246X.2009.04097.x>

Nissen, E., Elliott, J. R., Sloan, R. A., Craig, T. J., Funning, G. J., Hutko, A., et al. (2016). Limitations of rupture forecasting exposed by instantaneously triggered earthquake doublet. *Nature Geoscience*, 9(4), 330–336.

Ouillon, G., Ducorbier, C., & Sornette, D. (2008). Automatic reconstruction of fault networks from seismicity catalogs: Three-dimensional optimal anisotropic dynamic clustering. *Journal of Geophysical Research: Solid Earth*, 113(B1). doi:10.1029/2007JB005032.

Ouillon, G. & Sornette, D., (2011). Segmentation of fault networks determined from spatial clustering of earthquakes, *Journal of Geophysical Research: Solid Earth*, 116(B2). doi:10.1029/2010JB007752.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11), 559-572. <https://doi.org/10.1080/14786440109462720>

Perrin, C., I. Manighetti, J.-P. Ampuero, F. Cappa, and Y. Gaudemer (2016), Location of largest earthquake slip and fast rupture controlled by along-strike change in fault structural maturity due to fault growth, *Journal of Geophysical Research: Solid Earth*, 121, 3666–3685, <https://doi.org/10.1002/2015JB012671>

Petersen, G.M., Niemz, P., Cesca, S., Mouslopoulou, V., & Bocchini, G.M. (2021). Clusty, the waveform-based network similarity clustering toolbox: concept and application to image complex faulting offshore Zakynthos (Greece). *Geophysical Journal International*, 224(3), 2044-2059. <https://doi.org/10.1093/gji/ggaa568>

Piegari, E., Herrmann, M., & Marzocchi, W. (2022). 3-D spatial cluster analysis of seismic sequences through density-based algorithms. *Geophysical Journal International*, 230(3), 2073-2088. <https://doi.org/10.1093/gji/ggac160>

Piegari, E., Camanni, G., Mercurio, M., & Marzocchi, W. (2023a). A hypocenter clustering workflow to illuminate segmented fault surfaces. [Dataset]. Zenodo. <https://zenodo.org/record/8210903>

Piegari, E., Camanni, G., Mercurio, M., & Marzocchi, W. (2023b). hypo_clustering: August 3, 2023 Release (v1.0). [Software]. Zenodo. <https://zenodo.org/record/8211032>

Quinn, D. P., & Ehlmann, B. L. (2019). A PCA-based framework for determining remotely sensed geological surface orientations and their statistical quality. *Earth and Space Science*, 6(8), 1378-1408. <https://doi.org/10.1029/2018ea000416>

Roche, V., Camanni, G., Childs, C., Manzocchi, T., Walsh, J., Conneally, J., ... & Delogkos, E. (2021). Variability in the three-dimensional geometry of segmented normal fault surfaces. *Earth-Science Reviews*, 216, 103523. <https://doi.org/10.1016/j.earscirev.2021.103523>

Ross, Z. E., Cochran, E. S., Trugman, D. T., & Smith, J. D. (2020). 3D fault architecture controls the dynamism of earthquake swarms. *Science*, 368 (6497), 1357-1361. doi:10.1126/science.abb0779.

Truttmann, S., Diehl, T., & Herwegh, M. (2023). Hypocenter-based 3D Imaging of Active Faults: Method and Applications in the Southwestern Swiss Alps. *Journal of Geophysical Research: Solid Earth*, e2023JB026352. <https://doi.org/10.1029/2023JB026352>

Valoroso, L., Chiaraluce, L., Piccinini, D., Stefano, R., Schaff, D., Waldhauser, F., (2020). The 2009 Mw 6.1 L'Aquila normal fault system imaged by 64,051 high-precision foreshock and aftershock locations (Version 1.0). Zenodo. <https://doi.org/10.5281/zenodo.4036248>.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17, 395-416.

Walsh, J. J., & Watterson, J. (1989). Displacement gradients on fault surfaces. *Journal of Structural Geology*, 11(3), 307-316. [https://doi.org/10.1016/0191-8141\(89\)90070-9](https://doi.org/10.1016/0191-8141(89)90070-9)

Walsh, J. J., Bailey, W. R., Childs, C., Nicol, A., & Bonson, C. G. (2003). Formation of segmented normal faults: a 3-D perspective. *Journal of Structural Geology*, 25(8), 1251-1262. [https://doi.org/10.1016/S0191-8141\(02\)00161-X](https://doi.org/10.1016/S0191-8141(02)00161-X)

Wang, Y., Ouillon, G., Woessner, J., Sornette, D., & Husen, S. (2013). Automatic reconstruction of fault networks from seismicity catalogs including location uncertainty. *Journal of Geophysical Research: Solid Earth*, 118(11), 5956-5975. <https://doi.org/10.1002/2013JB010164>

Wang, H., Liu, M., Duan, B., & Cao, J. (2020). Rupture Propagation along Stepovers of Strike-Slip Faults: Effects of Initial Stress and Fault Geometry. *Bulletin of the Seismological Society of America*. 110 (3), 1011-1024. <https://doi.org/10.1785/0120190233>

Wesnousky, S. (1988). Seismological and structural evolution of strike-slip faults. *Nature* 335, 340–343. <https://doi.org/10.1038/335340a0>

Wesnousky, S. (2006). Predicting the endpoints of earthquake ruptures. *Nature* 444, 358-360. <https://doi.org/10.1038/nature05275>

Woessner, J., Laurentiu, D., Giardini, D., Crowley, H., Cotton, F., Grünthal, G., Valensise, G., Arvidsson, R., Basili, R., Demircioglu, M.B., Hiemer, S., Meletti, C., Musson, R.W., Rovida, A. N., Sesetyan, K., Stucchi, M., & The SHARE Consortium (2015). The 2013 European Seismic Hazard Model: key components and results. *Bull. Earthq. Eng.* 13, 3553-3596. <https://doi:10.1007/s10518-015-9795-1>

Wu, Y. M., Chang, C. H., Zhao, L., Teng, T. L., & Nakamura, M. (2008). A comprehensive relocation of earthquakes in Taiwan from 1991 to 2005. *Bulletin of the Seismological Society of America*, 98(3), 1471-1481. <https://doi.org/10.1785/0120070166>

Wu, Y. M., Zhao, L., Chang, C. H., & Hsu, Y. J. (2008). Focal-mechanism determination in Taiwan by genetic algorithm. *Bulletin of the Seismological Society of America*, 98(2), 651-661. <https://doi.org/10.1785/0120070115>

Yıkılmaz, M.B., Turcotte, D.L., Heien, E.M., Kellogg, L.H., & Rundle, J.B. (2015). Critical Jump Distance for Propagating Earthquake Ruptures Across Step-Overs. *Pure Appl. Geophys.* 172, 2195-2201. <https://doi.org/10.1007/s00024-014-0786-y>