J Seismol (2020) 24:443–457
https://doi.org/10.1007/s10950-020-09921-8

ORIGINAL ARTICLE

Seiscloud, a tool for density-based seismicity clustering and visualization

Simone Cesca

Received: 30 October 2019 / Accepted: 19 April 2020 / Published online: 7 May 2020

Abstract Clustering algorithms can be applied to seismic catalogs to automatically classify earthquakes upon the similarity of their attributes, in order to extract information on seismicity processes and faulting patterns out of large seismic datasets. We describe here a Python open-source software for density-based clustering of seismicity named seiscloud, based on the pyrocko library for seismology. Seiscloud is a tool to dig data out of large local, regional, or global seismic catalogs and to automatically recognize seismicity clusters, characterized by similar features, such as epicentral or hypocentral locations, origin times, focal mechanisms, or moment tensors. Alternatively, the code can rely on user-provided distance matrices to identify clusters of events sharing indirect features, such as similar waveforms. The code can either process local seismic catalogs or download selected subsets of seismic catalogs, accessing different global seismicity catalog providers, perform the seismic clustering over different steps in a flexible, easily adaptable approach, and provide results in form of declustered seismic catalogs and a number of illustrative figures. Here, the algorithm usage is explained and discussed through an application to Northern Chile seismicity.

Keywords Seismicity · Clustering · Location · Moment tensor

Highlights
- Open source Python tool for density-based clustering and visualization tools
- Easy processing of reference seismic catalogs or own datasets
- Flexible clustering upon different spatial, temporal of focal mechanism based metrics or using distance matrices

S. Cesca (✉)
Section 2.1 Physics of Earthquakes and Volcanoes, GFZ German Research Centre for Geosciences Potsdam, Helmholtzstr. 7, 14467 Potsdam, Germany
e-mail: simone.cesca@gfz-potsdam.de

1 Introduction

In recent years, the global densification of seismic stations, the growing interest in microseismicity monitoring, with the deployment of dense local networks to identify natural and anthropogenic microseismicity, and the implementation of powerful and unsupervised algorithms to scan large seismic datasets have allowed seismologists to detect, locate, and characterize increasingly weak (micro)earthquakes. As a consequence, seismic catalogs, reporting the most relevant earthquake parameters, are also becoming increasingly large. While large datasets could potentially provide more accurate insights into local, regional, and global seismic processes, digging the most relevant information out of large catalogs becomes challenging due to their growing size. Clustering algorithms are useful tools to automatically identify families of similar items out of large datasets: applied to seismicity, they can be used to detect earthquakes with similar features, such as hypocentral locations, origin times, magnitudes, or focal mechanisms. Such type of seismicity classification is important to support seismic monitoring programs and seismicity interpretation studies. For example, the application of a

temporal seismicity clustering, aimed at the identification of seismicity bursts occurring within short time frames, can be used to identify seismic sequences and swarms or, in the frame of monitoring issues, to early detect anomalous seismicity rates, potentially revealing stress or pore pressure transients, anticipating larger earthquakes (e.g., in the presence of foreshock activity) or volcanic unrests (e.g., if the seismicity is the result of magma/fluid propagation).

Seismicity clustering has been extensively used in the past years, mostly in the frame of single studies, for a variety of purposes, with a quite broad literature. Here, selected works are cited with the aim to provide an overview of the broad range of potential approaches and applications. Spatial clustering, either based on the spatial distribution of epicenters, hypocenters, and/or centroids, is probably the most simple and used approach. Aiming at identifying confined seismogenic regions, their shape, orientation and extent, spatial clustering methods have been used to compare cluster geometries with potentially active faults (Ansari et al. 2009), to image complex fault networks (Ouillon and Sournette 2011), to reconstruct earthquake ruptures over multiple fault segments (Cesca et al. 2017) and in the frame of seismicity forecasting (Lippiello et al. 2012). Schoenball and Ellsworth (2018) used a density-based spatial clustering approach to identify induced seismicity clusters and active faults and to track the evolution of the moment release over individual faults. A comparative approach, discussing the results of different spatial clustering techniques applied to a common seismogenic region, was provided by Konstantaras et al. (2012).

Temporal and spatio-temporal clustering has been used in first order to identify aftershocks and to decluster seismic catalogs, to investigate the temporal evolution of seismicity and the earthquake recurrence times over different time periods and depth intervals (e.g., Kagan and Knopoff 1976; Reasenberg 1985, Frohlich 1987; Kagan and Jackson 1991), for spatio-temporal cluster identification (Schaefer et al. 2017), to discuss statistical models for triggered seismicity (Hainzl et al. 2000; Sornette and Werner 2005), to identify earthquake sequences in regions with relevant swarm activity (Jacobs et al. 2013), and to detect and image growing fractures (Maghsoudi et al. 2014).

Focal mechanism clustering aims at the separation of families of earthquakes presenting different rupture styles and orientations. Mostly based on the pioneering work by Kagan (1991), who introduced a commonly accepted definition for the similarity of double-couple (DC) earthquake focal mechanisms, through the angle which bears his name. Focal mechanism-based clustering algorithms have been used to identify different families of mining-induced (Cesca et al. 2014) and natural earthquakes in different seismotectonic contexts (Cesca et al. 2016, 2017; Custódio et al. 2016) as well as their temporal evolution in response to anthropogenic and earthquake-induced stress perturbations.

Waveform-based clustering approaches are nowadays the most in vogue. The formulation basically links the earthquake similarity to the similarity among earthquakes' direct observations, in the form of single or multiple recorded waveforms (Maurer and Deichmann 1995; Cattaneo et al. 1999; Moriya et al. 2003; Wehling-Benatelli et al. 2013; Cesca et al. 2020). The observation of similar waveforms for different earthquakes implies a similarity among their location, depth, and focal mechanism. In this sense, waveform-based clustering approaches are more strict than the previously formulated ones, and identified clusters are composed by very similar events, such as earthquake repeaters (see Uchida and Bürgmann 2019 for a review). Following the opposite formulation, so-called waveform template approaches use the recorded waveforms or waveform features of pre-identified earthquakes to search for similar, weaker signals within continuous data streams (e.g., Yoon et al. 2015).

An open question remains how to combine clustering results based on different earthquake parameters. In Custódio et al. (2016), independent clustering based on hypocentral location and focal mechanisms have been used to reconstruct a complex pattern of faulting in a region of diffused seismicity. An alternative approach, aiming at combining different metrics into a common framework, has been formulated by Lasocki (2014) and tested on a mining-induced dataset (Lizurek and Lasocki 2014). Joint information on location, time, and magnitude of earthquakes have been used by Zaliapin et al. (2008) to formulate an approach to identify aftershocks, and later by Zaliapin and Ben Zion (2013) to detected earthquake clusters in California.

This paper describes the seiscloud software, implementing a density-based clustering approach for seismicity. The algorithm has a flexible implementation, with several different metrics available, offering to cluster seismicity upon the similarity of the spatial, temporal, or focal mechanism earthquake parameters, or to use user-defined metrics by providing distance matrices.

Seiscloud can process own seismic catalogs, with different extent of available information (including, when available additional information such as magnitude, focal mechanism or full moment tensors), or download seismic catalogs from open global datasets. Seismicity clustering results can be easily visualized, through the generation of a number of illustrative figures. The software is implemented in Python3, open source, it uses and requires the installation of the pyrocko libraries and it is available through an open git platform (https://git.pyrocko.org/cesca/seiscloud).

2 Density-based clustering for seismicity

Density-based clustering algorithms scan the dataset to be clustered, searching for densely populated regions. If the items density overcomes a user-defined threshold, a cluster is identified, and its edges will be defined where the item density drops below the given threshold. In terms of seismicity, the concept is most easily explained considering the spatial distribution of earthquake hypocenters as a metric. Within seismogenic volumes, e.g., at plate boundaries, the density of seismic foci is typically high, whereas this decreases with distance from the plate boundary. Therefore, one or more seismicity clusters can be found there, as the conditions are easily met to have a sufficient number of neighboring hypocenters.

Among the many different clustering techniques, we rely here on DBSCAN (Ester et al. 1996), as one of the most used density-based clustering algorithms. The performance of DBSCAN is controlled by two parameters controlling the density threshold: one defines the minimum number of neighboring items (*Nmin*) and one the maximum acceptable distance (ε). The condition to create a cluster is that there exist an item i and at least *Nmin* other items j with a distance $d_{i,j} < \varepsilon$. In this case, a cluster is formed. Item i is then defined as *core* item, while items j are defined as *density-reachable* items from the *core* item i. Each of the *density-reachable* items is then investigated to check if it also lies in a densely populated region: if they have a sufficient number of neighbors, they are also *core* items, while if they lie in lower density regions, they are referred as *edge* items. Both *core* and *edge* items will be assigned to a cluster. Besides *core* and *edge* items, we also have *isolated* items, which are neither located in densely populated regions, nor are *density-reachable* from a *core* item: *isolated* items are basically located in low density regions and will not be assigned to any cluster. Figure 1 illustrates the difference among *core*, *density-reachable*, and *isolated* items (earthquakes) with a simple, spatial clustering seismicity application.

This procedure illustrates some of the specific features of DBSCAN and, in general, of density-based clustering methods. First of all, the result of the clustering is not dependent on the items sorting, as this does not affect the identification of *core*, *edge*, and *isolated* items. Second, the handling of *isolated* items implies the recognition of the concept of outliers: basically, the clustering algorithm does not force all items to be clustered, and items which do not respect the clustering conditions are simply treated as outliers, or unclustered items. Another important aspect, with relevant implications for seismicity clustering, is related to the concept of density-reachability: if item b is reachable from item a, and item c is reachable from item b, this means that c is *density-reachable* from item a and they will pertain to a common cluster. However, items a and c are not necessarily similar among them. Using again the spatial distribution of hypocentral location as example, two far distant earthquakes may be allocated in a common cluster if there is a highly active seismogenic region which connects the hypocenters of the two earthquakes. Furthermore, the dense region path connecting two items can have a very peculiar or curved shape; thus, spatially clustered earthquakes do not necessarily map a single, planar fault, but could also concatenate earthquakes distributed along adjacent fault segments and/or complex fault systems, if these are well connected. The choice of the density parameters *Nmin* and ε controls the result of the clustering procedure. While this could be seen as a method drawback, the flexible definition of the two parameters has the advantage to allow the user for different grade of clustering resolution, which can extract and highlight information with a variable accuracy (Cesca et al. 2014). Figure 2 shows different results of seismicity clustering obtained for a common seismic catalog, when selecting two different setups of the clustering parameters. Basically, the selection of the *Nmin* and ε parameters controls the number, size, and heterogeneity of the resolved clusters, as well as the fraction of unclustered items. Increasing *Nmin* will require a larger number of neighbors: this tends to increase the size of the resolved clusters and generally to reduce their number, as small clusters are either merged or lost. The parameter ε controls the similarity constraint: reducing its value typically forces the clusters to be more

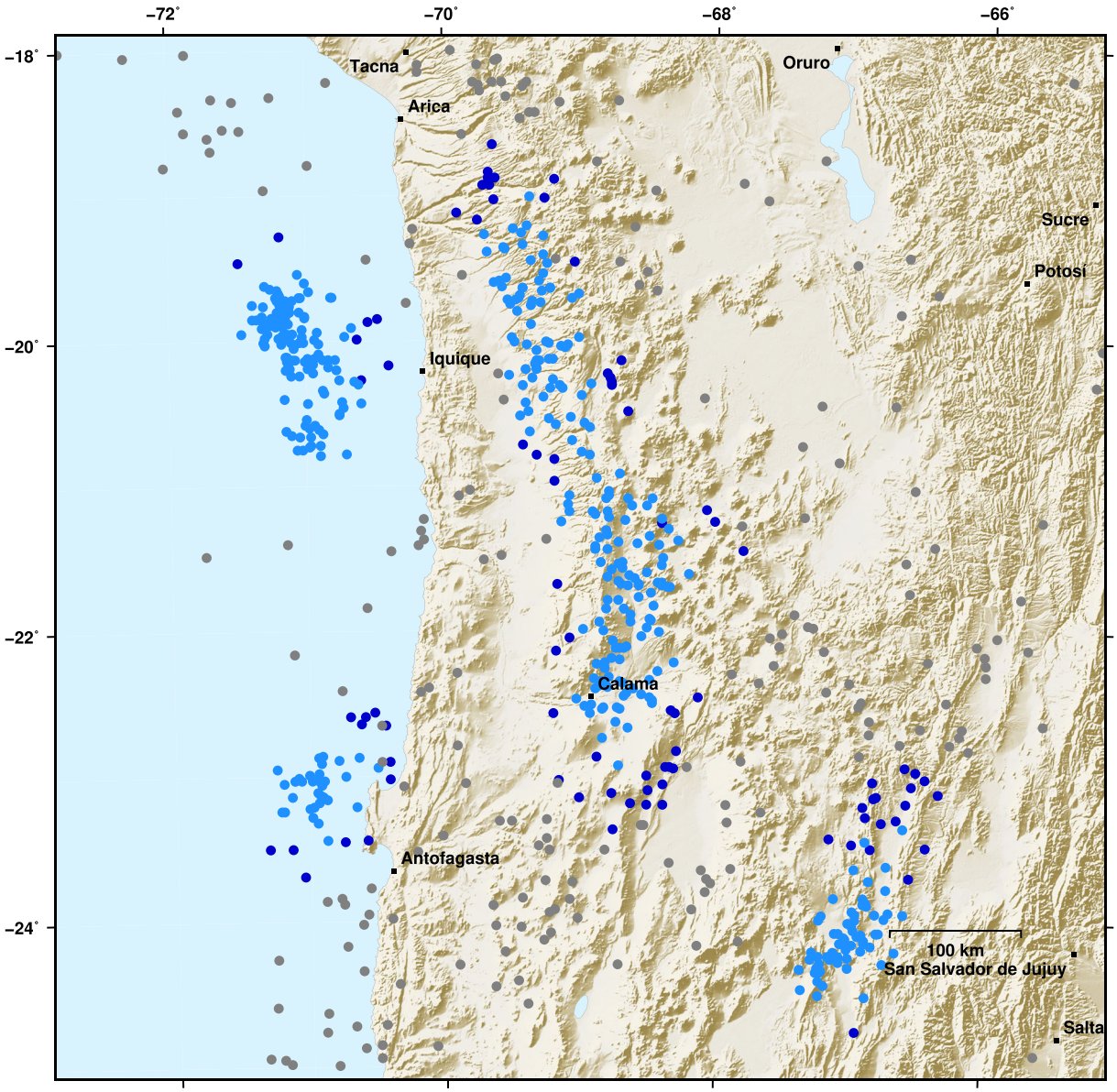

Fig. 1 Core (light blue circles), edges (dark blue circles), and isolated (gray circles) hypocenters as obtained applying the seiscloud algorithm for a spatial clustering application (*Nmin* = 20, $\varepsilon = 0.05$) to seismicity in Northern Chile (based on the Global CMT seismic catalog for the time period January 1, 1989–January 1, 2019 with the following spatial constraints: latitudes − 25° to − 18° N, longitudes − 73° to 67° East, depths from 0 to 170 km)

homogeneous (i.e. composed of very similar items), while increasing it will allow for more heterogeneous clusters. Consequently, the choice of ε will also affect the number and size of the clusters: for a common *Nmin*, small ε values typically lead to identify a larger number of smaller, more homogeneous clusters, and large ε values to less, larger, heterogeneous clusters. Another parameter to take into account is the number of the unclustered items, i.e., those items which do not correspond to any cluster: the fraction of unclustered events typically increases by reducing ε, because fewer item pairs will fulfill the smaller distance constraint, and by increasing *Nmin*, because fewer clusters will fulfill the larger cluster size constraint.

It is difficult to provide a single, general hint on how to select the DBSCAN parameters, as clustering algorithms could be used for different purposes and the same algorithm, applied with different setups to a common

dataset, could dig out different information. The choice of the parameters will thus need to necessarily take into account the seismicity features to resolve. Reducing as possible the number of unclustered events is often a desired feature, as the clustering results will be representative for the vast majority of the considered earthquakes. This result, favored by lower *Nmin* and higher ε values, may specifically benefit those applications which aim at recognizing general, first order seismicity patterns from a seismic catalog. On the contrary, if the goal is the detection of small homogeneous features, the choice should be driven to the resolution of small, compact clusters, by reducing *Nmin* and ε values.

Two additional, important considerations should be done on the choice of the DBSCAN parameters. First, it is worth noting that *Nmin* has a direct dependency on the size of the seismic catalog. Let us consider a steady seismogenic region, with similar seismicity rate, spatial distribution, or focal mechanisms over time. Then, the size of the seismic catalog would linearly increase with the duration of the considered time period: in these conditions, applying the seismic clustering with common DBSCAN parameters will provide different results if we analyze, e.g., 1 or 10 years of data. This implies that the choice of the *Nmin* value should be chosen according to the size of the catalog, eventually as a ratio of the seismic catalog size; in the previous conditions, we should increase by a factor N the value of *Nmin*, when we process a catalog N times larger. As for the choice of ε, this should be done taking into account the distance uncertainty, which depends on the uncertainty of seismic source parameters. If the location (or the focal mechanism) accuracy is poor, very high similarity among hypocentral locations (or very low Kagan angle among the focal mechanisms) should be avoided. Similarly, if seismic data are noisy, a high waveform similarity should not be required, because even for perfect repeaters, their theoretically highly similar waveforms will be noise contaminated and their correlation coefficient well below 1.0. In conclusion, ε should be chosen well above the limit imposed by the average distance uncertainty.

3 Metrics

In seiscloud, the DBSCAN implementation relies on normalized distance values in the range (0, 1), with distance 0 meaning equal items, and distance 1 denoting very different items (see later discussion). Dealing with seismicity, different metrics are proposed to assess the similarity or dissimilarity of earthquakes. Each of these metrics is then normalized to account values in the interval (0.0, 1.0). For some metrics, where the distance is defined over a finite interval, the normalization is unique, for example, when using the waveform correlation as a measure of earthquake similarity, a cross-correlation equal to 1.0 and − 1.0 (perfectly correlating and anti-correlating waveforms) will map into normalized distances of 0.0 and 1.0, respectively. In other cases, where a maximal distance cannot be defined (for example, when considering the spatial or time difference among two earthquakes), we will assign the maximal normalized distance 1.0 to all spatial or temporal differences, which are equal or larger than a given threshold (e.g., default thresholds are 1000 km for spatial distances and 365 days for the temporal distance).

Considering that most seismological studies rely on a point source approximation, neglecting the spatial and temporal finiteness of the earthquake rupture and only considering a spatial location (hypocenter or centroid) and time (origin or centroid time), the basic spatio-temporal description of an earthquake source is typically given by 4 parameters only: latitude, longitude, depth, and time. These are the basic earthquake attributes listed in any seismic catalog, from local applications to a global scale. In these conditions, spatial metrics, i.e., estimating the Euclidean distance among hypocenters or epicenters (in the case the earthquake depth is unavailable or unreliable) are obvious candidates to define a distance among two earthquakes. Similarly, the inter-event time, i.e., the absolute time among origin times, will describe the temporal similarity of two earthquakes. As discussed before, neither of these spatial and temporal metrics is defined in the interval (0, 1), nor can they be uniquely normalized. However, considering global applications as the largest potential targets for spatial clustering, and the duration of historical records in seismic catalogs for temporal clustering, we define 1000 km as the maximum distance of interest among two earthquakes and 1 year as the maximum inter-event time (these values can be easily modified, as described in the seiscloud manual). Thus spatial and temporal distances are divided by these normalization coefficients and any value exceeding 1.0 is then replaced by the maximum normalized distance (1.0).

Definition of metrics in the space of double couple (DC) focal mechanisms and moment tensor (MT)

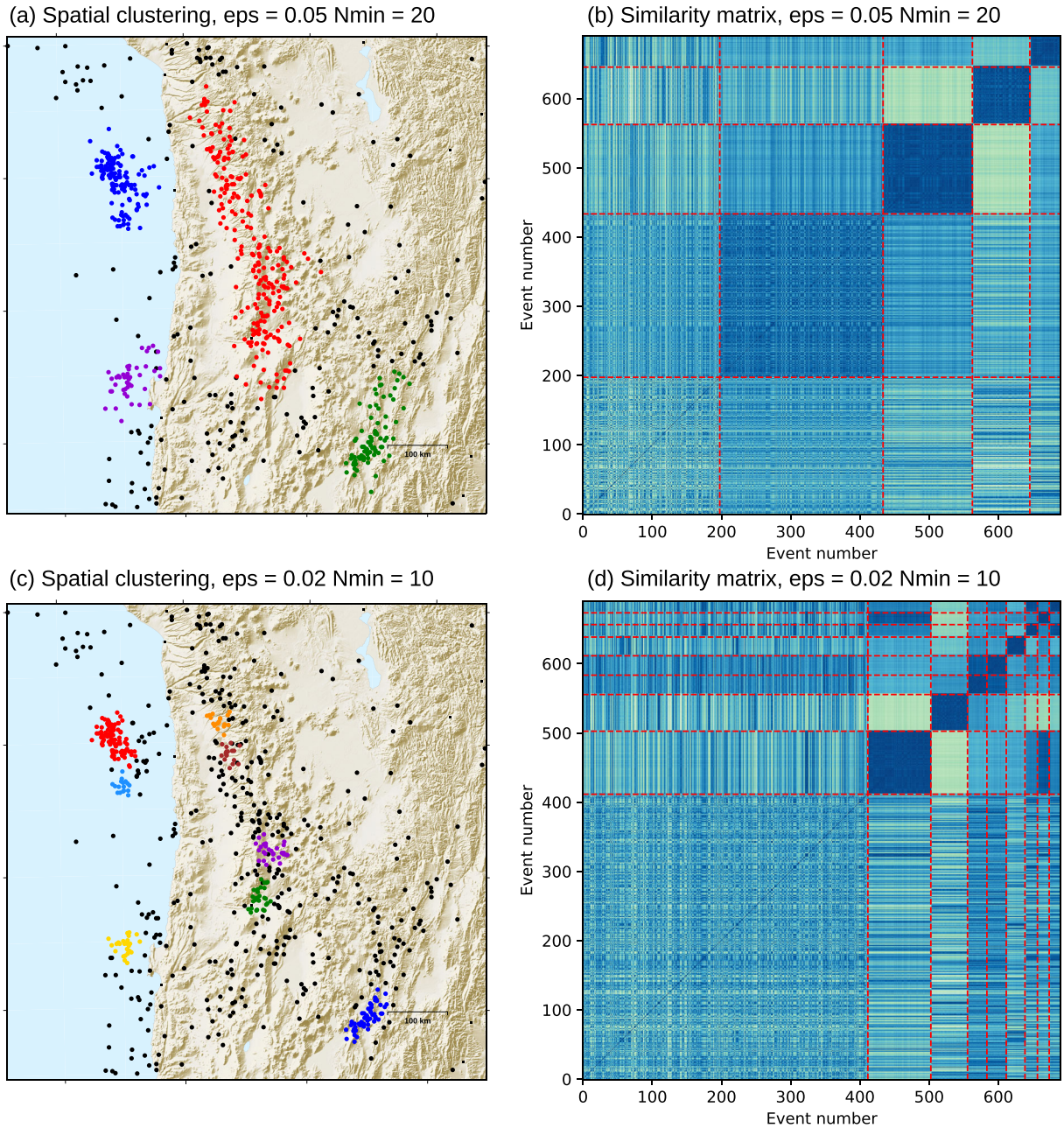

Fig. 2 Different spatial clustering results are obtained from the same catalog (Global CMT, latitudes 18–25° S, longitudes 73–65° W, depths 0–170 km, time span January 1, 1989–January 1, 2019) using different *Nmin* and ε parameters. The setup with *Nmin* = 20 and $\varepsilon = 0.05$ identifies four large and less homogeneous clusters (top, panel showing the map of clusters (**a**) and panel showing the similarity matrix sorted after clustering (**b**), blue colors denoting similar hypocentral locations, with unclustered and four clusters identified by red lines). The setup with *Nmin* = 10 and $\varepsilon = 0.02$ identifies many small and very homogeneous clusters (bottom, panel showing the map of clusters (**c**) and panel showing the similarity matrix sorted after clustering (**d**))

representations of the earthquake source have been proposed by Kagan (1991, 1992), Willemann (1993), and Tape and Tape (2012). In seiscloud, the Kagan angle (normalized by its maximum value of 120°), which is defined as the rotation angle to transform the DC focal mechanism of one earthquake into the one of the second earthquake, is used as a metric to assess the similarity among focal mechanism pairs. As for moment tensors,

following the metrics recompilation and discussion in Cesca et al. (2014), seiscloud offers a number of normalized metrics based on the moment tensor components, with variable fixed and flexible weighting; the adoption of flexible weighting, in particular, is suggested for cases where the resolution of the moment tensor components is not homogeneous, e.g., in the case of very shallow earthquakes inverted using low frequency surface waves (Bukchin et al. 2010; Valentine and Trampert 2012). Figure 3 illustrates the difference among clustering results from the same dataset, when clustering upon different metrics, either based on centroid location, centroid time, or DC focal mechanism.

The proposed metrics are used to compute distance matrices, which are square matrices of the size of the seismic catalog, listing the value of the distance among each pair of earthquakes (*i, j*). Distance matrices have 0 values on the diagonal, where each event is compared with itself (note that this notation differs from the one adopted by similarity matrices, which are also often used in seismology, where the highest similarity on the matrix diagonal is filled by 1 values, corresponding to the highest correlation), and are symmetric, requiring that distances are invariant upon the earthquake sorting, $d(i,j) = d(j,i)$. Note that this condition has some implications for the temporal metrics, which requires absolute differential times lacking any information on the temporal order among two earthquakes, and, potentially, for other user-defined metrics. In fact, besides using internally defined metrics, users can simply supply own distance matrices, obtained by considering any wished (normalized) measure of the similarity among two earthquakes.

4 Sample application: seismicity at the northern Chile subduction

The seiscloud software is implemented in Python3 and makes use of a number of standard libraries, including numpy (Oliphant 2006). Additionally, it requires the previous installation of the pyrocko Python library for seismology (Heimann et al. 2017), for seismic source and seismic catalog handling, and GMT version 5 (Wessel et al. 2013), for plotting issues. The seiscloud software is open source and can be downloaded via git at https://git.pyrocko.org/cesca/seiscloud, where additional information and a technical manual are also available.

As an example of application, we will consider the northern Chile subduction region, in the volume confined within latitude − 25° N and − 18° N, longitude − 72° E, and − 67° E and depths between 0 and 700 km. We consider 30 years of data from 1 January 1989 until 1 January 2019 and a broad magnitude range between Mw magnitude 4 and 10. During this time period, the region hosted a number of large Mw > 7.5 earthquakes and seismic sequences in Antofagasta (Mw 8.0, 1995; Ruegg et al. 1996; Delouis et al. 1997; Sobiesiak 2000), Tarapaca (Mw 7.8, 2005; Delouis and Legrand 2007), Tocopilla (Mw 7.7, 2007; Delouis et al. 2009; Peyrat et al. 2010; Schurr et al. 2012), and Iquique (Mw 8.2 and Mw 7.7 largest aftershock, 2014; Schurr et al. 2014; Ruiz et al. 2014; Kato et al. 2016; Cesca et al. 2016).

Seiscloud can be called alone or with a suffix *–help* to show a basic command help. Typically the seiscloud call is accompanied by one of the following subcommands: *example*, *init*, *matrix*, *cluster*, and *plot*, to perform the different steps of the clustering procedure. All calls with subcommands can be accompanied by a suffix *–help*, to provide a basic help on the subcommand usage, and/or with the suffix *–force* to force some file/directory overwrite.

As a first step, a seiscloud configuration file has to be created. The command:

seiscloud example

provides a first example of configuration file (*seiscloud.example.config*). The configuration file defines the spatial, temporal, and magnitude targets. Besides processing a local seismic catalog, seiscloud allows for the query of open global catalogs, such as the one provided by Global CMT (Dziewonski et al. 1981; Ekström et al. 2012), which is used for this example and contains 604 events for the selected spatial, temporal, and magnitude intervals. The seiscloud configuration file should also list the basic clustering configuration, including the chosen metric and the values of the clustering parameters *Nmin* and ε. Since the Global CMT catalog provides information on the moment tensor, we apply here a moment tensor-based metrics, namely the Kagan angle among DC components. DBSCAN parameters are fixed to $Nmin = 20$ and $\varepsilon = 0.1$, which implies that a cluster will be created whenever for one target earthquake, there are at least 20 other events with a DC mechanism sufficiently similar to the DC of the target one; the similarity threshold ε of 0.1 corresponds to a Kagan angle of 12°. The used seiscloud configuration file is hereafter called *seiscloud.example.config*, using

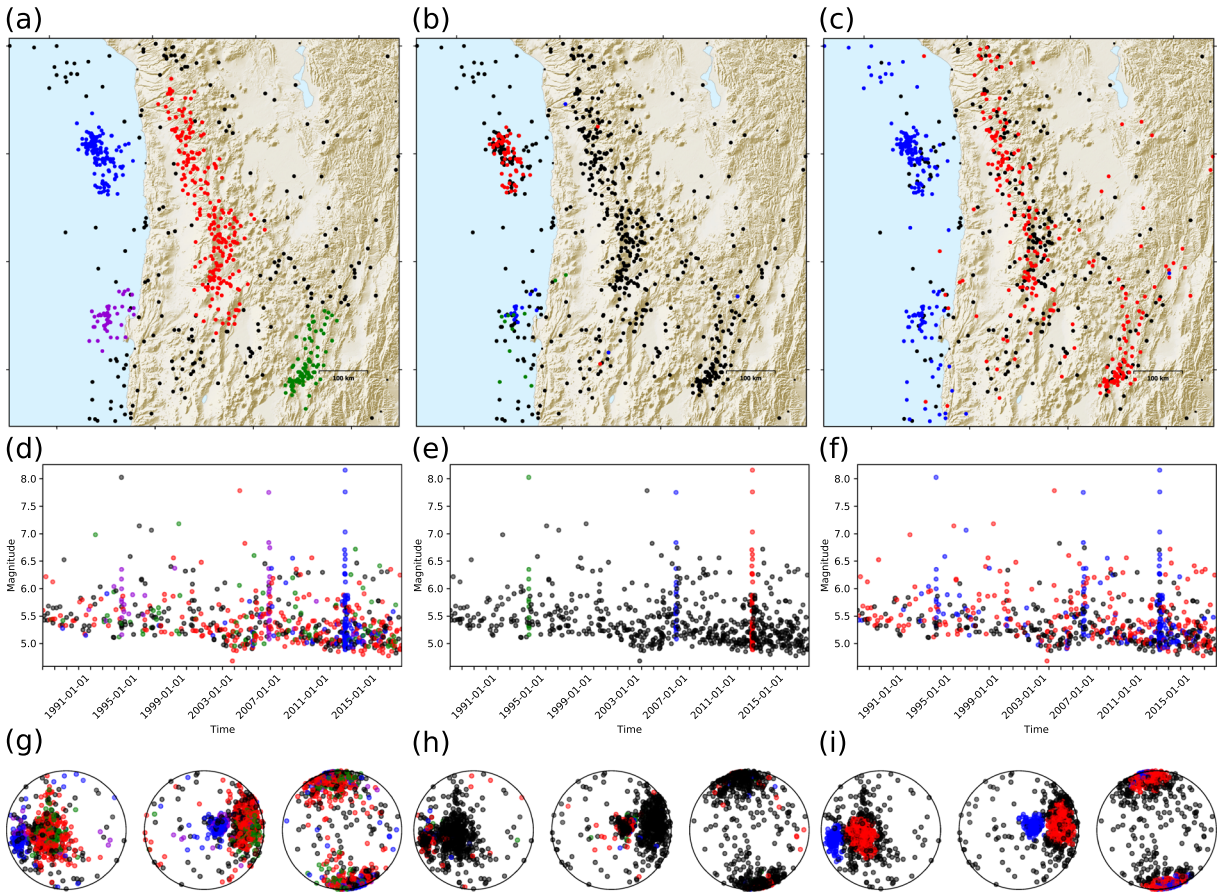

Fig. 3 Comparison of clustering using spatial (left), temporal (center), and focal mechanism (right) metrics for a common catalog (Global CMT, latitudes 18–25° S, longitudes 73–65° W, depths 0–170 km, time January 1, 1989–January 1, 2019). The spatial clustering was performed with *Nmin* = 20 and $\varepsilon = 0.05$ (at least 20 earthquakes within 50 km), the temporal clustering with *Nmin* = 10 and $\varepsilon = 0.02$ (at least 10 earthquakes within 7.3 days), and the double couple focal mechanism clustering with *Nmin* = 25 and $\varepsilon = 0.1$ (at least 25 earthquakes with Kagan angle below 12°). **a**, **b**, **c** plots show hypocentral locations in map view; **d**, **e**, and **f,** the time evolution of magnitudes; and **g**, **h**, **i**, the distribution of DC principle axes (pressure, tension, and null axis). Colors correspond to clusters identified in each approach (red and blue for the two largest clusters, other colors, when present, for smaller clusters, and black for unclustered events). The spatial clustering identifies four clusters: two shallow ones, detecting the Iquique (blue) and Antofagasta and Tocopilla (purple) sequences, and two with intermediate depth (red, green) seismicity. The temporal clustering identifies the largest sequences in the region: Iquique 2014 (red), Tocopilla 2007 (blue), and Antofagasta 1995 (green). The focal mechanism clustering identifies two main families: thrust parallel to the plate margin at shallow depths (blue) and normal faulting with the same orientation at intermediate depths (red)

for simplicity the name which is automatically generated.

Running the command:

seiscloud init seiscloud.example.config

creates a project directory, including copies of the catalog and configuration files.

The computation of a similarity matrix is performed with the command:

seiscloud matrix seiscloud.example.config

where a suffix *–view* opens a graphical window to visualize the distance matrix and a suffix *–savefig* saves a copy of the distance matrix plot in the project directory. The palette is blue to white, for highest to lowest earthquake similarity, respectively. Note that the distance matrix computation depends on the metric but not on the clustering parameters *Nmin* and ε.

Once the distance matrix is computed, the density-based clustering can be performed with the command:

seiscloud cluster seiscloud.example.config

which result now depends on the choice of *Nmin* and ε. This step can be repeatedly run, changing the clustering parameters, with no need to recompute the distance

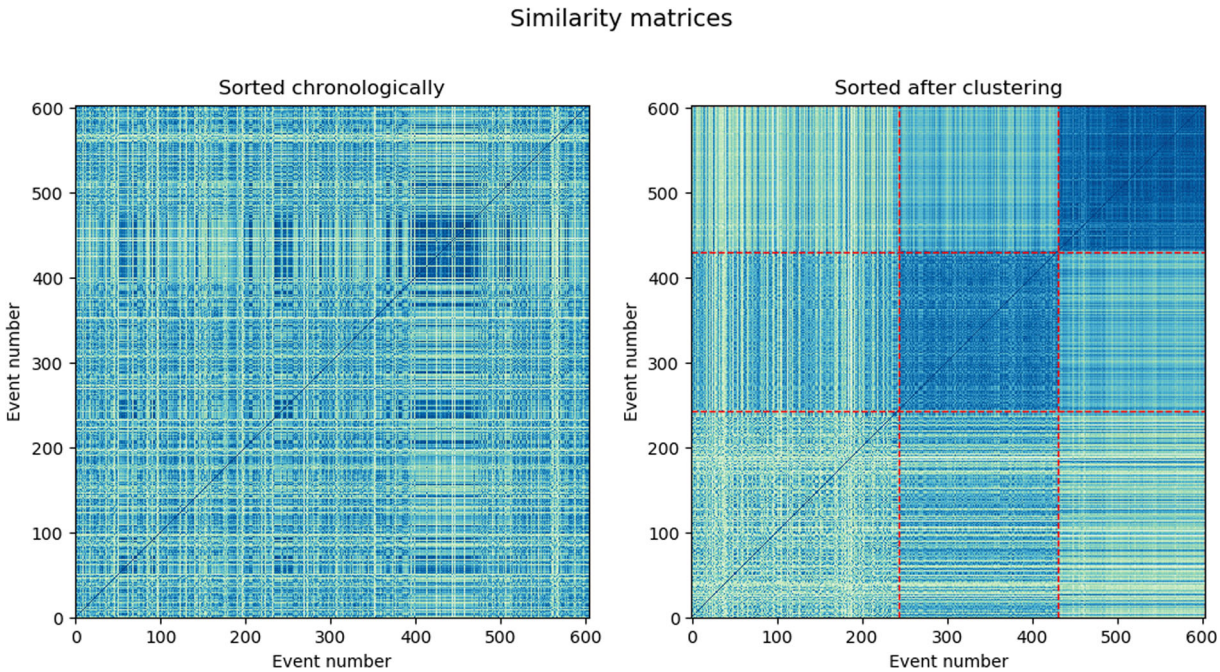

Fig. 4 Similarity matrices obtained using the example configuration file (Global CMT, latitudes − 25° to − 21° N, longitudes − 72° to − 67° E, depths 0 to 700 km, time January 1, 1989–January 1, 2019,) before (left) and after (right) clustering. Blue denotes the highest similarity, white lowest similarity. Dashed red lines on the similarity matrix after clustering helps to visualize the different clusters: in this application, the first 244 events are not clustered, the following first cluster has 187 events, and the second and last cluster has 173 events

matrix. Also the cluster subcommand supports the options –view and –savefig, analogous to the previous step, but now showing the distance matrices before and after the clustering (with earthquakes sorted by cluster, Fig. 4). The command creates a subdirectory *clustering_results* and stores there clustered seismic catalogs.

Finally, the command:

seiscloud plot seiscloud.example.config

creates a subdirectory *clustering_plots* and saves there 11 illustrative plots. Some of these are shown in the following figures for the sample application. A first plot shows the distance matrices (Fig. 4) before (i.e., with earthquakes sorted in chronological order) and after the clustering (where events are sorted upon their cluster). Three plots are dedicated to the spatial distribution of seismicity: one shows the epicentral distribution colored according to the type of clustering items (i.e., core, edge, and isolated earthquakes), one colored according to the clusters and one including focal mechanisms (when available) and colored according to the cluster, (Fig. 5). Two plots describe the temporal evolution of seismicity (magnitude and depth as function of time, Fig. 6). The geometry of DC focal mechanism is illustrated by a plot of the distribution of their principle axis(Fig. 7), one with the median mechanisms for each cluster (Custódio et al. 2016) and one of a triangle diagram representation (after Frohlich 1992), which illustrates the dominance of strike-slip, normal, and thrust faulting components. Non-DC components of full MT solutions are shown with a Hudson plot (Hudson et al. 1989), which is helpful to appreciate the sign and contribution of isotropic and compensated linear vector (CLVD) components against the DC component. Finally, one last plot describes the results of the clustering in the frame of a normalized time-space distance plot (after Zaliapin et al. 2008). Focal sphere and dots referring to single events in all these plots (with the exception of the map with the type of clustering items) are colored according to the clustering: unclustered events are plotted in black, events in single clusters are colored, sorted by the cluster size (by default, as in this work, the largest cluster is plotted in red, then blue, green, dark violet, gold, dark orange, and so on, further details are provided

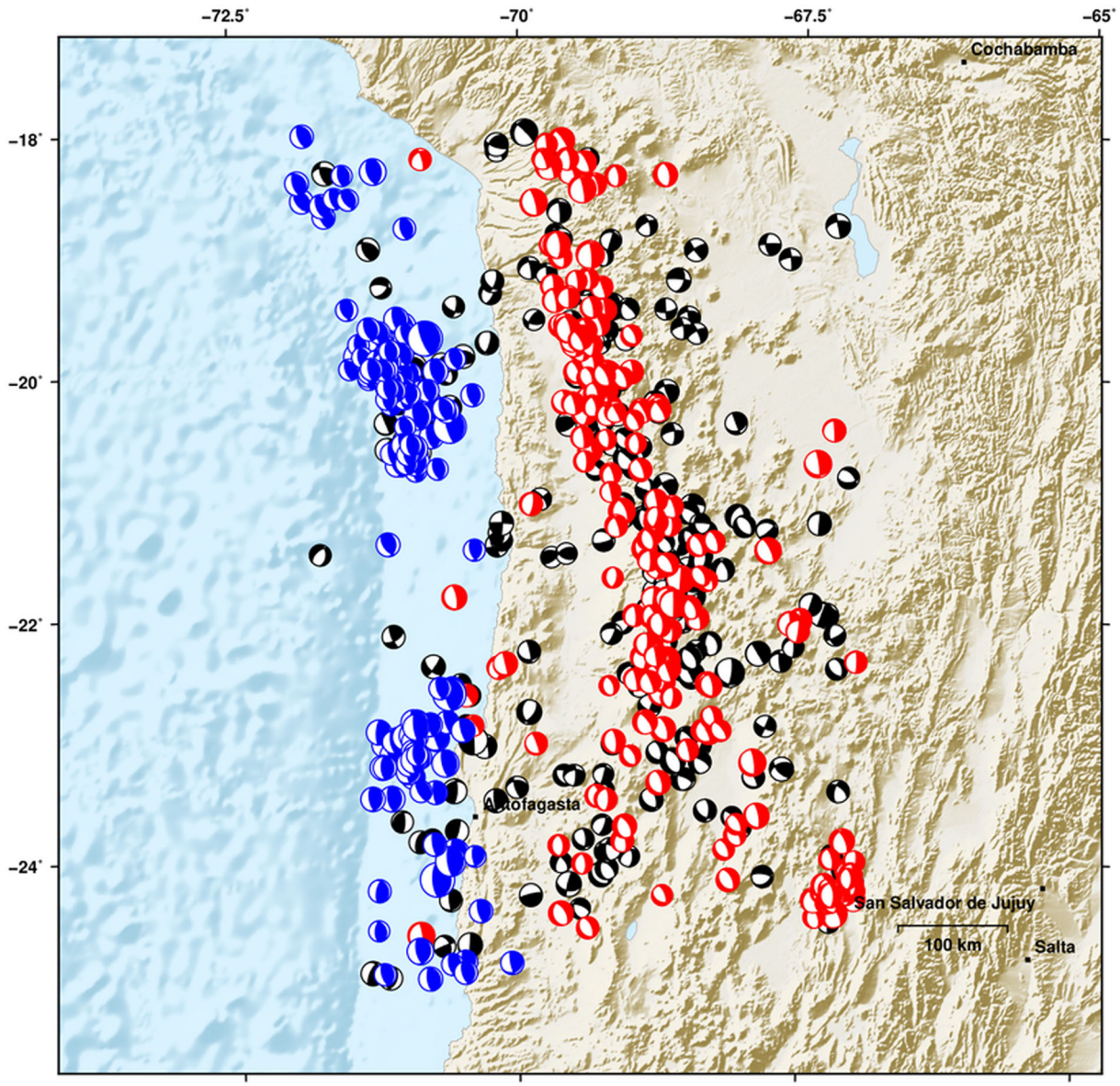

Fig. 5 Clustering results showing the spatial distribution of epicenters and focal mechanisms. The largest cluster (red) is composed by similarly oriented thrust mechanisms, associated to the Chilean subduction. The second cluster (blue) is composed of intermediate depth normal faulting events, typical of the deeper segment of the subduction between 70 and 350 km depth. Unclustered events (black) are characterized by a variety of different focal mechanisms

in the seiscloud manual online, https://git.pyrocko.org/cesca/seiscloud/src/master/gx/seiscloud_manual.pdf).

5 Discussion and conclusions

This work describes a Python-based software, seiscloud, for the clustering and visualization of seismic catalogs. The algorithm can process local catalogs, including as minimum information epicentral locations and origin time. When available, additional information, such as depth, magnitude, focal mechanism, and/or moment tensor, can be included and can be used for seismicity clustering. Alternatively, seiscloud can access remote databases to download seismic catalogs for a selected region, allowing for filtering the time, location, depth, and magnitude intervals. The clustering procedure,

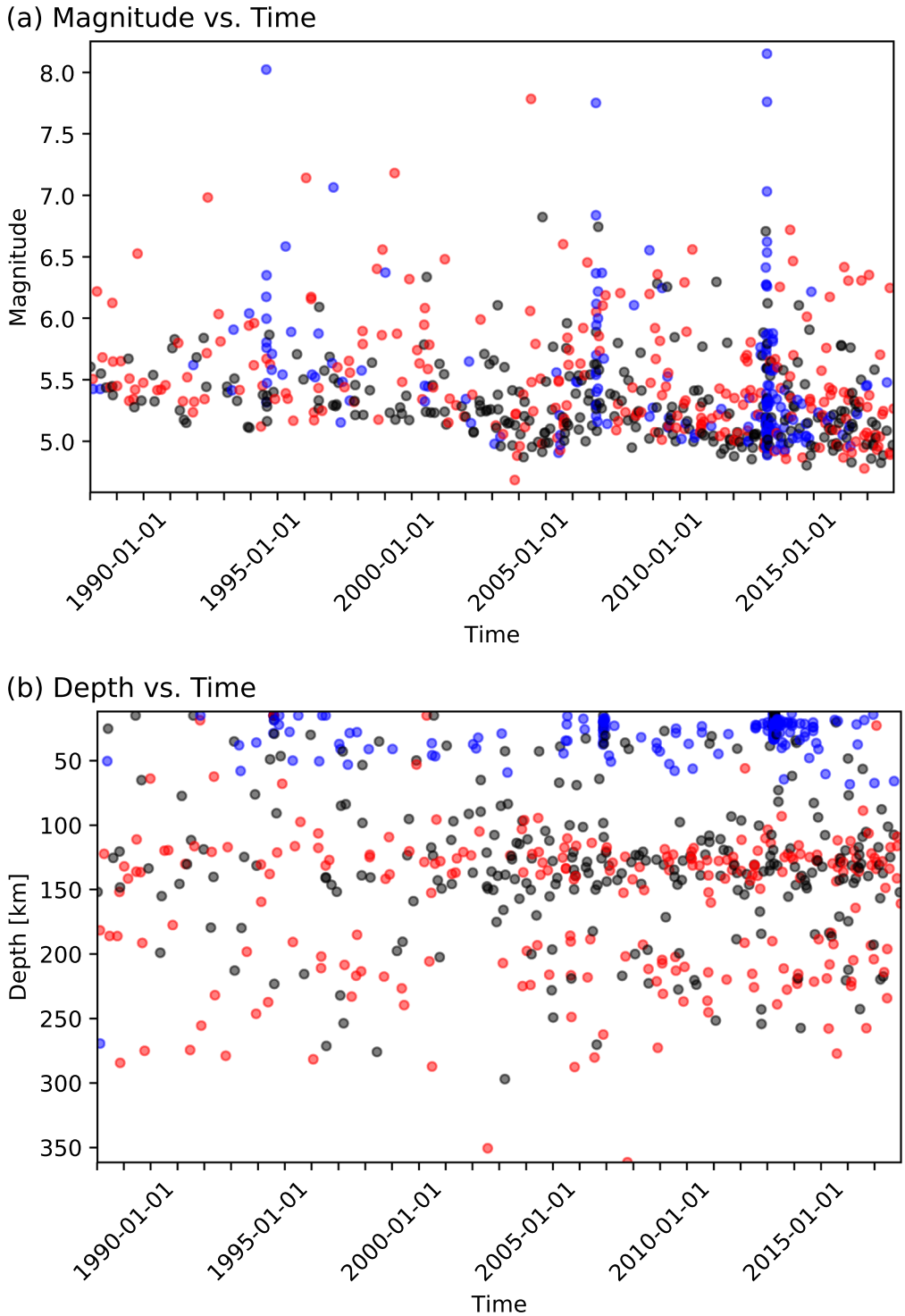

Fig. 6 Clustering results showing the temporal evolution of seismicity in form of magnitude versus time (**a**) and depth versus time (**b**). Largest earthquakes often occur as thrust mechanisms at shallow depth, with the exception of the 2005 Tarapaca earthquake, which is deeper and normal faulting. The temporal evolution highlights the major seismic sequences accompanying large thrust subduction earthquakes: Antofagasta (1995), Tocopilla (2007), and Iquique (2014). Events are plotted with the same color scale as in Fig. 5

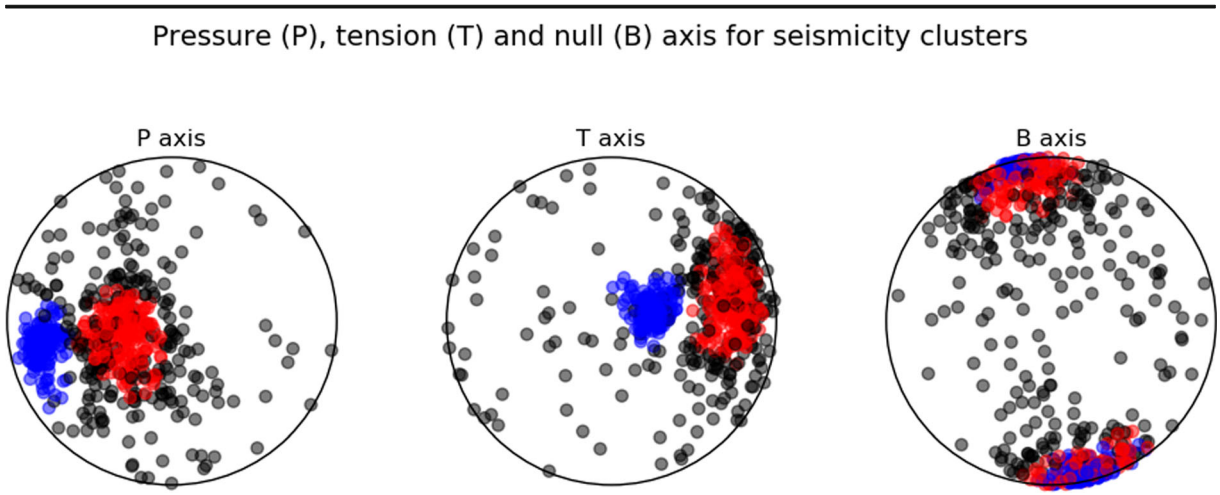

Fig. 7 Clustering results showing the distribution of pressure (P), tension (T), and null (B) axis of the deviatoric moment tensor solutions. The difference among thrust (red) and normal (blue) faulting is here visualized by the rotation of pressure and tension axis, while the null axis remains approximately parallel to the plate margin. Events are plotted with the same color scale as in Fig. 5

based on the implementation of the DBSCAN clustering, can be performed to consider the similarity of locations, times, focal mechanisms, or moment tensors, using a number of predefined metrics. Alternatively, users can provide own similarity matrices, based on different analyses, such as waveform correlation.

While clustering tools have been proposed and used in the past, only few are freely available (e.g., CLUSTER2000, Reasenberg 1985; FMC, Álvarez-Gómez 2019) and typically limited to few metrics. Most of the studies dealing with seismicity clustering approaches, including those implementing density-based algorithms, are not accompanied by open dedicated software. In this context, a major improvement introduced by seiscloud is its support of many different metrics, which are not limited to consider the earthquake location, time and magnitude. A second, substantial advantage of seiscloud lies in the combination of the clustering routines with visualization scripts, which are automatically generating high quality, helpful figures, useful for the better understanding and visualization of the clustering results and supporting results interpretation.

By providing multiple measures of the earthquake similarity, seiscloud can also support the future development and testing of multi-dimensional clustering approaches, where the similarity among different earthquake attributes may be considered. This type of analysis remains to date at a relatively early stage, with few tested approaches (e.g., Zaliapin et al. 2008; Lasocki 2014; Custódio et al. 2016), but could be supported in the future in the frame of automated, unsupervised processing able to handle big datasets. One approach for multi-dimensional clustering can be based on the implementation of single, joint metrics, simultaneously accounting for the similarity of different parameters. This approach can already be implemented by seiscloud, provided a proper distance matrix is given. Currently proposed approaches (e.g., Lasocki 2014) rely on the distribution of source parameter values to define metrics in multi-parameters spaces. While this approach offers a simple metric definition, a drawback could be that the distance between two earthquakes does not only depend on their source parameters, but also on the parameter distributions of the catalog. In this way, extending the catalog to broader times or to broader regions could change the distance value for the same earthquake pair. Another problem is that such implementation neglects the different uncertainties on different parameters. The second approach, tested, e.g., by Custódio et al. (2016) to define seismic clusters in the Western Mediterranean and Atlantic, requires the parallel clustering using different earthquake attributes, and the a posteriori combination of independent clustering results for single attributes. While the single clustering procedures can be performed with seiscloud, their integration needs to be performed separately.

Seiscloud can be used to explore and visualize large seismic catalogs and for a variety of applications, such as the prompt identification of anomalous seismicity, with localized hypocenters, or earthquake swarms, to map active fault geometry and to reconstruct complex faulting by mapping regions with similar locations and

focal mechanisms, and to track microseismicity patterns induced by anthropogenic operations, such as fluid injection, mining or water reservoir operations. Preliminary versions of the algorithms were applied in the last years for different studies in tectonic, volcanic, and induced seismicity environments, illustrating a number of potential applications: identification of spatial clusters at diffuse plate margins (Custódio et al. 2016), reconstruction of complex rupture faults from aftershock distributions (Cesca et al. 2017), detection and growth of fractures in mining regions upon fluid injection (López-Comino et al. 2017) and thermal perturbations (Maghsoudi et al. 2014), identification and characterization of clusters of earthquakes with similar focal mechanisms for tectonic (Cesca et al. 2016; Custódio et al. 2016) and anthropogenic (Cesca et al. 2014) seismicity, and also identification of seismic sources radiating similar seismic signals at volcanoes (Gaete et al. 2019; Cesca et al. 2020). Seiscloud is open source and can be found at https://git.pyrocko.org/cesca/seiscloud, where additional technical information is available.

Acknowledgments The author is thankful to the editor, Prof. Mariano Garcia-Fernandez, and an anonymous reviewer, as well as to Dr. S. Heimann and the pyrocko developers team, for useful discussions and to support the pyrocko git repository hosting seiscloud.

Funding Information Open Access funding provided by Projekt DEAL.

Data availability Data used in this study (basically, an extended seismic catalog) is available at the Global CMT (Dziewonski et al. 1981; Ekström et al. 2012) website (https://www.globalcmt.org/) and can be automatically downloaded by seiscloud. The target region of the study case has been extensively monitored in the frame of the Integrated Plate boundary Observatory Chile (IPOC) project (http://www.ipoc-network.org/).

References

Álvarez-Gómez J (2019) FMC—earthquake focal mechanisms data management, cluster and classification. SoftwareX 9: 299–307. https://doi.org/10.1016/j.softx.2019.03.008

Ansari A, Noorzad A, Zafarani H (2009) Clustering analysis of the seismic catalog of Iran. Comput Geosci 35:475–486. https://doi.org/10.1016/j.cageo.2008.01.010

Bukchin B, Clévédé E, Mostinskiy A (2010) Uncertainty of moment tensor determination from surface wave analysis for shallow earthquakes. J Seismol 14:601–614. https://doi.org/10.1007/s10950-009-9185-8

Cattaneo M, Augliera P, Spallarossa D, Lanza V (1999) A waveform similarity approach to investigate seismicity patterns. Nat Hazards 19:123–138

Cesca S, Şen AT, Dahm T (2014) Seismicity monitoring by cluster analysis of moment tensors. Geophys J Int 196:1813–1826. https://doi.org/10.1093/gji/ggt492

Cesca S, Grigoli F, Heimann S, Dahm T, Kriegerowski M, Sobiesiak M, Tassara C, Olcay M (2016) The Mw 8.1 2014 Iquique, Chile, seismic sequence: a tale of foreshocks and aftershocks. Geophys J Int 204:766–1780. https://doi.org/10.1093/gji/ggv544

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

Cesca S, Letort J, Razafindrakoto HNT, Heimann S, Rivalta E, Isken MP, Nikkhoo M, Passarelli L, Petersen G, Cotton F, Dahm T (2020) Drainage of a deep magma reservoir near Mayotte inferred from seismicity and deformation. Nat Geosci 13(1):87–93. https://doi.org/10.1038/s41561-019-0505-5

Custódio S, Lima V, Vales D, Cesca S, Carrilho F (2016) Imaging active faulting in a region of distributed deformation from the joint clustering of focal mechanisms and hypocentres: application to the Azores–western Mediterranean region. Tectonophysics 676:70–89. https://doi.org/10.1016/j.tecto.2016.03.013

Delouis B, Legrand D (2007) Mw 7.8 Tarapaca intermediate depth earthquake of 13 June 2005 (northern Chile): fault plane identification and slip distribution by waveform inversion. Geophys Res Lett 34. https://doi.org/10.1029/2006GL028193

Delouis B, Monfret T, Dorbath L, Pardo M, Rivera L, Comte D, Haessler H, Caminade JP, Ponce L, Kausel E, Cisternas A (1997) The Mw= 8.0 Antofagasta (northern Chile) earthquake of 30 July 1995: a precursor to the end of the large 1877 gap. Bull Seismol Soc Am 87:427–445

Delouis B, Pardo M, Legrand D, Monfret T (2009) The M w 7.7 Tocopilla earthquake of 14 November 2007 at the southern edge of the northern Chile seismic gap: rupture in the deep part of the coupled plate interface. Bull Seismol Soc Am 99: 87–94

Dziewonski AM, Chou TA, Woodhouse JH (1981) Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. J Geophys Res 86:2825–2852. https://doi.org/10.1029/JB086iB04p02825

Ekström G, Nettles M, Dziewonski AM (2012) The global CMT project 2004-2010: centroid-moment tensors for 13,017 earthquakes. Phys Earth Planet Inter 200-201:1–9. https://doi.org/10.1016/j.pepi.2012.04.002

Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis E, Han J, Fayyad UM (eds) Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press, pp 226–231

Frohlich C (1987) Aftershocks and temporal clustering of deep earthquakes. J Geophys Res 92:13944–13956

Frohlich (1992) Triangle diagrams: ternary graphs to display similarity and diversity of earthquake focal mechanisms. Physics Earth Planet Int 75(1–3):193–198. https://doi.org/10.1016/0031-9201(92)90130-N

Gaete A, Cesca S, Franco L, San Martin J, Cartes C, Walter TR (2019) Seismic activity during the 2013–2015 intereruptive phase at Lascar volcano, Chile. Geophys J Int 219:449–463. https://doi.org/10.1093/gji/ggz297

Hainzl S, Zöller G, Kurths J (2000) Self-organization of spatio-temporal of earthquake clusters. Nonlin Process Geophys 7: 21–29

Heimann S, Kriegerowski M, Isken M, Cesca S, Daout S, Grigoli F, Juretzek C, Megies T, Nooshiri N, Steinberg A, Sudhaus H, Vasyura-Bathke H, Willey T, Dahm T (2017) Pyrocko-an open-source seismology toolbox and library. GFZ Data Services. https://doi.org/10.5880/GFZ.2.1.2017.001

Hudson JA, Pearce RG, Rogers RM (1989) Source type plot for inversion of the moment tensor. J Geophys Res 94:765–774

Jacobs KM, Smith EGC, Savage MK, Zhuang J (2013) Cumulative rate analysis (CURATE): a clustering algorithm for swarm dominated catalogs. J Geophys Res Solid Earth 118:553–569. https://doi.org/10.1029/2012JB009222

Kagan YY (1991) 3-D rotation of double-couple earthquake sources. Geophys J Int 106:709–716

Kagan YY (1992) Correlation of earthquake focal mechanisms. Geophys J Int 110:305–320

Kagan YY, Jackson DD (1991) Long-term earthquake clustering. Geophys J Int 104:117–133

Kagan YY, Knopoff L (1976) Statistical search for non-random features of the seismicity of strong earthquakes. Phys Earth planet Int 12:291–318

Kato A, Fukuda J, Kumazawa T, Nakagawa S (2016) Accelerated nucleation of the 2014 Iquique, Chile MW 8.2 earthquake. Sci Rep. https://doi.org/10.1038/srep24792

Konstantaras AJ, Katsifarakis E, Maravelaiks E, Skounakis E, Kokkinos E, Karapidakis E (2012) Intelligent spatial-clustering of seismicity in the vicinity of the Hellenic seismic arc. Earth Sci Res 1. https://doi.org/10.5539/esr.v1n2p1

Lasocki S (2014) Transformation to equivalent dimensions—a new methodology to study earthquake clustering. Geophys J Int 197:1224–1235

Lippiello E, Marzocchi W, de Arcangelis L, Godano C (2012) Spatial organization of foreshocks as a tool to forecast large earthquakes. Sci Rep 2. https://doi.org/10.1038/srep00846

Lizurek G, Lasocki S (2014) Clustering of mining-induced seismic events in equivalent dimension spaces. J Seismol 18:543–563

López-Comino JA, Cesca S, Heimann S, Grigoli F, Milkereit C, Dahm T, Zang A (2017) Characterization of hydraulic fractures growth during the Äspö Hard Rock Laboratory experiment (Sweden). Rock Mech Rock Eng 50:2985–3001

Maghsoudi S, Hainzl S, Cesca S, Dahm T, Kaiser D (2014) Identification and characterization of growing large-scale en-echelon fractures in a salt mine. Geophys J Int 196: 1092–1105. https://doi.org/10.1093/gji/ggt443

Maurer H, Deichmann N (1995) Microearthquake cluster detection based on waveform similarities, with an application to the western Swiss Alps. Geophys J Int 123:588–600

Moriya H, Niitsuma H, Baria R (2003) Multiplet-clustering analysis reveals structural details within the seismic cloud at the Soultz geothermal field, France. Bull Seismol Soc Am 93: 1606–1620. https://doi.org/10.1785/0120020072

Oliphant TE (2006) A guide to NumPy (Vol. 1). Trelgol Publishing USA

Ouillon G, Sournette D (2011) Segmentation of fault networks determined from spatial clustering of earthquakes. J Geophys Res 116. https://doi.org/10.1029/2010JB007752

Peyrat S, Madariaga R, Buforn E, Campos J, Asch G, Vilotte JP (2010) Kinematic rupture process of the 2007 Tocopilla earthquake and its main aftershocks from teleseismic and strong-motion data. Geophys J Int 182:1411–1430

Reasenberg P (1985) 2nd-order moment of Central California seismicity, 1969–1982. J Geophys Res Solid Earth and Planets 90:5479–5495

Ruegg JC, Campos J, Armijo R, Barrientos S, Briole P, Thiele R, Arancibia M, Cañuta J, Duquesno T, Chang M, Lazo D, Lyon-Caen H, Ortlieb L, Rossignol JC, Serrurier L (1996) The Mw= 8.1 Antofagasta (North Chile) earthquake of July 30, 1995: first results from teleseismic and geodetic data. Geophys Res Lett 23:917–920. https://doi.org/10.1029/96GL01026

Ruiz S, Metois M, Fuenzalida A, Ruiz J, Leyton F, Grandin R, Vigny C, Madariaga R, Campos J (2014) Intense foreshocks and a slow slip event preceded the 2014 Iquique Mw 8.1 earthquake. Science 345:1165–1169. https://doi.org/10.1126/science.1256074

Schaefer AM, Daniell JE, Wenzel F (2017) The smart cluster method. J Seismol 21:965–985. https://doi.org/10.1007/s10950-017-9646-4

Schoenball M, Ellsworth W (2018) A systematic assessment of the spatiotemporal evolution of fault activation through induced seismicity in Oklahoma and southern Kansas. J Geophys Res Solid Earth 122:10,189–10,206. https://doi.org/10.1002/2017JB014850

Schurr B, Asch G, Rosenau M, Wang R, Oncken O, Barrientos S, Salazar P, Vilotte JP (2012) The 2007 M7. 7 Tocopilla northern Chile earthquake sequence: implications for along strike and downdip rupture segmentation and megathrust frictional behavior. J Geophys Res Solid Earth. https://doi.org/10.1029/2011JB009030

Schurr B, Asch G, Hainzl S, Bedford J, Hoechner A, Palo M, Wang R, Moreno M, Bartsch M, Zhang Y, Oncken O, Tilmann F, Dahm T, Victor P, Barrientos S, Vilotte J (2014) Gradual unlocking of plate boundary controlled initiation of the 2014 Iquique earthquake. Nature 512:299–302. https://doi.org/10.1038/nature13681

Sobiesiak MM (2000) Fault plane structure of the Antofagasta, Chile earthquake of 1995. Geophys Res Lett 27:577–580

Sornette D, Werner MJ (2005) Apparent clustering and apparent background earthquakes biased by undetected seismicity. J Geophys Res 110. https://doi.org/10.1029/2005JB003621

Tape W, Tape C (2012) Angle between principal axis triples. Geophys J Int 191:813–831

Uchida N, Bürgmann R (2019) Repeating earthquakes. Annu Rev Earth Planet Sci 47:305–332. https://doi.org/10.1146/annurev-earth-053018-060119

Valentine AP, Trampert J (2012) Assessing the uncertainties on seismic source parameters: towards realistic error estimates for centroid-moment-tensor determinations. Phys Earth Planet Int 210:36–49. https://doi.org/10.1016/j.pepi.2012.08.003

Wehling-Benatelli S, Becker D, Bischoff M, Friederich W, Meier T (2013) Indications for different types of brittle failure due to active coal mining using waveform similarities of induced seismic events. Solid Earth Discuss 5(1):655–698

Wessel P, Smith WHF, Scharroo R, Luis J, Wobbe F (2013) Generic mapping tools: improved version released. EOS Trans AGU 94:409–410. https://doi.org/10.1002/2013EO450001

Willemann RJ (1993) Cluster analysis of seismic moment tensor orientations. Geophys J Int 115:617–634

Yoon CE, O'Reilly O, Bergen KJ, Beroza GC (2015) Earthquake detection through computationally efficient similarity search. Sci Adv 1(11):e1501057

Zaliapin I, Ben Zion Y (2013) Earthquake clusters in southern California I: identification and stability. J Geophys Res 118(6):2847–2864. https://doi.org/10.1002/jgrb.50178

Zaliapin I, Gabrielov A, Keilis-Borok V, Wong H (2008) Clustering analysis of seismicity and aftershock identification. Phys Rev Lett 101:018501